Interferometric Methods

James Christopher Kent

Cavendish Laboratory
&
Wolfson College

UNIVERSITY OF
CAMBRIDGE

July 2020

A dissertation submitted for the degree of Doctor of Philosophy

Summary

Future radio telescopes promise great advances in resolution and sensitivity. These include the Square Kilometer Array, a two array instrument, in South Africa and Australia. Similarly, the next generation Very Large Array (ngVLA) is being designed for construction in North America. These arrays all promise exceptional advances in sensitivity, angular resolution, and survey speed. The SKA and ngVLA are both specified to have sensitivities at the level of μJy's. The SKA-Low instrument will consist of a huge number of dipoles antennas in Australia which is pushing the bounds of current FX correlator technology with $\mathbb{O}(n^2)$ scaling, where n is the number of antennas. The design proposals for these instruments include a dense core of antennas, necessitating advances in imaging methods for these very dense cores versus more traditionally sparse instruments.

Another ambitious experiment is the Hydrogen Epoch of Reionisation Array (HERA) in South Africa which hopes to make the first direct detection of the Epoch of Reionisation through the red-shifted HI signal which is a factor of 10^5 smaller than the thermal-like noise.

In this thesis, these problems are tackled by re-examining the underlying principles of interferometry. The first working example of a direct imaging correlator is presented which allows images to be formed directly from the voltages off each antenna in a dense array, without an expensive cross-correlation operation as is typically required. A detailed discussion is given of how standard steps in interferometric imaging differ in this new scheme, including calibration. Additionally the first wide field direct imaging correlator is presented, which allows the problems of non-coplanarity to be dealt with for both sparse and dense arrays in a very efficient manner on modern GPU compute hardware. These are, to the best of the authors knowledge, the only working implementations of a direct imaging correlator for generic arrays with no restrictions on the geometry of the array or homogeneity of constituent receiver elements. These new approaches have been published in the scientific literature as discussed in the Declaration.

Moving on from this, the closure phase bispectrum is presented as a way of uncovering the cosmological Epoch of Reionisation signal from the HI line. This is using the HERA telescope, which consists of a dense core of parabolic antennas in a highly redundant layout. A data

reduction and processing pipeline for the HERA telescope is constructed and presented, for use with the bispectrum. Initial results towards a cosmologial limit are reported.

The HERA telescope relies on redundancy in its antenna elements for its calibration and measurement strategy. The bispectrum with its unique mathematical propeties, in combination with forward modelling, is shown to be a potent tool for probing departures from the assumed reudundancy. It is shown, through this method, that HERA suffers significant direction-dependent non-redundancies in the dataset used for our analysis, which are extremely difficult to calibrate out.

Finally, the problem of wide-field imaging in next generation arrays is tackled through the development and implementation of a new scheme of wide field imaging. This uses a new method of parallelising the problem of wide-field imaging, and is intended for use with the very large datasets that will be produced by upcoming instruments. Two schemes are introduced: w-towers, and Improved w-towers. The latter generalises the former in combination with advances in optimal convolution theory for the radio astronomy "gridding" problem. The theory behind this approach is explored, and a high performance implementation is presented for w-towers and Improved w-stacking within Improved w-towers.

Contents

Declaration

As required, I declare that this thesis is the result of my own work. Any work or results arising from collaboration are clearly explained in the text. This thesis is the result of work completed between October 2016 and September 2019 in the Astrophysics group of the Cavendish Laboratory, with a period of studying abroad at Arizona State University from July 2018 to September 2018. This thesis has not been submitted for an award or qualification at any other university, and does not exceed 6×10^4 words in length.

The work presented in Chapter 5 was done in collaboration with Dr Bojan Nikolic, Dr Chris Carilli and Dr Nithyanandan Thyagarajan, of the University of Cambridge and the National Radio Astronomy Observatory. The results presented however are my own independent analysis.

The work presented in Chapter 6 was done in collaboration with Dr Peter Wortmann of the Astrophysics group of the Cavendish Laboratory. Peter had the original idea for splitting a wide field imaging problem in this way, with high performance implementation work done by myself. I have also undertaken some mathematical proofs of some of the more interesting results, to convince others.

As specified in the chapters themselves, the results or methodologies presented in this thesis have been published in the following papers of which I am the principal author or co-author:

1) J. Kent et al., "A real-time, all-sky, high time resolution, direct imager for the long wavelength array", MNRAS, Volume 486, Issue 4, July 2019, Pages 5052-5060
2) J. Kent et al., "Direct wide-field radio imaging in real-time at high time resolution using antenna electric fields", MNRAS, Volume 491, Issue 1, January 2020, Pages 254-263
3) C. L. Carilli, N. Thyagarajan, J. Kent, B. Nikolic et al., "Imaging and Modelling Data from the Hydrogen Epoch of Reionization Array", Accepted by Astrophysical Journal Supplement, arXiv:2002.07692
4) N. Thyagarajan, C. L. Carilli, B. Nikolic, J. Kent et al., "Detection of Cosmic Structures using the Bispectrum Phase. II. First Results from Application to Cosmic Reionization Using the Epoch of Reionization Array", Accepted by Physical Review Letters D., arXiv:2005.10275

James Christopher Kent

Acknowledgements

I would like to thank my supervisors, Bojan Nikolic and Steve Gull, who have contributed greatly to my time here in the Cavendish Laboratory, making my experience educational and enjoyable. Bojan exposed me to many techniques in computing that I had not previously encountered, as well as presenting with me interesting opportunities for research that have proven very successful. He has always been very encouraging and his sharp wit and jovial nature has made my time in the Cavendish as much fun as it has been work. I have picked up his love of open source software, in particular GNU EMACS. To Steve I owe an enormous amount, his formidable mastery of mathematics and physics has opened my mind to an entirely different way of thinking about problems which has proved very fruitful. Together they have both ensured I have learnt at least a few things during my time here. I'd also like to thank Dave Green for his thesis LaTeX template. Thank you to ARM Ltd. for providing the funding to complete this research.

I thank also Peter Wortmann, for his encouragement and patient guidance as I grappled with the problems of wide-field imaging which we tackled together. His training as a computer scientist exposed me to an intuition on the problems of radio astronomy which is uncommon in the field, and has been very educational. His refreshing directness has led to some very productive and 'energetic' discussions. He has also regularly indulged my interest in functional progamming with HASKELL, which was his particular specialty from his own PhD, and also chess. He has tirelessly put up with my poor German.

From my time studying and working abroad in the USA during my PhD, I would like to thank Judd Bowman, Chris Carilli, Adam Beardsley, Danny Jacobs, Jayce Dowell, Gregg Taylor, and Nithyanandan Thyagarajan. I greatly enjoyed the research I did in Arizona and New Mexico, and you all provided excellent advice and guidance for my research and career at large which will serve me well for years to come. To Jayce I would like to extend a particular thanks for his excellently written software and documentation for the LWA, which along with his expertise and familiarity with the systems, made the development, implementation, and validation of the groundbreaking EPIC correlator a highly enjoyable experience.

I am fortunate to have had parents, Sarah and Adam, who have been encouraging of my

scientific interest since I was a young boy, enjoying my success and nodding politely when I tried to explain things to them about my research. There were always chemistry sets and science books around, and varying tolerance for my dismantling of household electronics to learn more about them, which certainly helped promote my interest in science and engineering.

Thank you to Sarah Cowans for being encouraging as I started this journey. I would also like to thank, in no particular order, Max, Sabrina, Sonny, Harry, George, Stephen, Felicity, Izzy, James L, Cameron, Holly and all my other friends for making the years spent studying in this beautiful city of Cambridge unforgettable. Finally to Roxane, thank you for the great conversation, good times, and cherished memories.

CHAPTER 1

RADIO INTERFEROMETRY AND APERTURE SYNTHESIS

1.1 Overview

Radio Interferometry is the usage of multiple radio antennas to form a measurement of the sky overhead. This is facilitated by the use of optical spatial coherency to form coherent images of the sky, limited by the characteristics of the instrument.

The ideas of astronomical radio interferometry were developed after the end of the Second World War, after the demobbing of English, Dutch, American and Australian scientists, who had developed expertise in radar during wartime research. The ideas of astronomical interferometry had been developed practically by Michelson in Michelson and Pease (1921), to make an optical angular measurement of α-Orionis, also known as Betelguese. This was by interfering two beams of partially coherent light and observing the amplitude of the interfered light.

This was taken further and formalised in Zernike (1938) , in what would become a landmark paper in optical theory, showing that an image could be formed from multiple spatially seperated measurements. The formalism is generic and applies to instruments of all frequencies, however there are significant instrumental difficulties in building imaging interferometers beyond the microwave domain.

This generic optical theory is the crux of radio interferometric imaging. This "Aperture Synthesis" allows multiple radio antennas dispersed across some distance to synthesise an aperture of a telescope with an angular resolution defined by the largest seperation between two

antenna elements in the array. Angular resolution can be taken to be $\theta_{res} \approx \frac{1.22\lambda}{d}$, where λ is the wavelength of the light, d the diameter of the array defined by the largest seperation between two antennas, or in the case of a single dish radio telescope by the diameter of the collecting dish. Thus an interferometer allows angular resolution to be achieved in a way that would not be possible with a single-dish telescope. The largest single dish telescopes in the world cannot match the angular resolution of a modestly sized interferometric array, of a similar frequency. However, single dish telescopes often have higher sensitivity due to a greater collecting area, as well as lacking the complex point spread function of an interferometer.

Whilst interferometric instruments do allow excellent angular resolution, in practice there is a significant number of other effects, mostly negative, that have to be carefully taken into consideration to allow a coherent image of the sky to be recovered. These can be loosely collected into the problems of calibration, imaging (particularly wide-field imaging), and deconvolution. This thesis predominantly looks at several questions in imaging, however calibration and deconvolution are still important, and discussed appropriately.

The problems of calibration revolve around solving for various contributing gains in the telescope and array, such as the amplitudes and phases of each antenna in the array, as well as correcting for phase terms incurred by the ionosphere. Ionospheric correction is particularly important for larger arrays and at higher frequencies. Additionally there has to be correction for the direction-dependent power pattern of the beam of each antenna. These problems can be resolved through a variety of different schemes as described in Thompson et al. (2017).

Deconvolution is another important step in the formation of a radio interferometric image, where an inverse problem is solved to correct for a true sky measurement from a dataset that has information of the true sky convolved with the optical point spread function of the instrument. In a radio interferometer this point spread function is also known as the 'dirty beam', and is often of a complex nature due to the incomplete sampling of the spatial coherence function.

This thesis revolves around new methods of interferometric data analysis, in particular new imaging schemes that have never been demonstrated before, that have attractive benefits compared to the current state of the art. This thesis also describes using higher order data products from an interferometer to make cosmological measurements as well as instrumental diagnostics from highly redundant arrays. Whilst the schemes are practical in their nature and implementation, there is a not insignificant element of theoretical work that has gone into them and is described within individual chapters.

In this vein, this introductory Chapter seeks to introduce various mathematical derivations and conventions that will be referred to throughout the thesis. In Section 1.2 the basic interferometric imaging equation is derived from first principles, and some observations made that will be expanded on in subsequent Chapter's. Some of the subtleties involved are elaborated

to help make subsequent points clearer. Much of this material was derived with the help of Thompson et al. (2017) and Born and Wolf (1999).

In Section 1.3, the details of the practical radio interferometer are elaborated. Some important observations and properties of the noise properties of the radio interferometer are presented, with a detailed derivation provided in Appendix A.

1.2 Aperture Synthesis Mapping

The following section sets out the important theories for Aperture synthesis mapping. First assume an interferometer of N antennas, with each antenna having some physical location in coordinates x, y, z which are defined on a set of orthonormal basis vectors e_1, e_2, e_3 in $\mathbb{R}^3$, with units of meters. Overhead, there is some sky brightness distribution $I(l, m, n)$, where l, m, n represent local (in the frame of the interferometer) direction cosine coordinates on the "celestial sphere". The celestial sphere is a 2-sphere, $\mathbb{S}^2$ in $\mathbb{R}^3$, based on the assumption that most astrophysical objects are in the far field and thus the radial component of the spherical coordinates of the celestial sphere can be neglected. n is degenerate with l and m, such that $n = \sqrt{1 - l^2 - m^2}$. It is natural, and useful, to define u, v, w coordinates that are in terms of wavelengths, λ, where $u = \frac{x}{\lambda}$, $v = \frac{y}{\lambda}$ and $w = \frac{z}{\lambda}$. Here λ represents the wavelength at a particular frequency sampled by interferometer.

1.2.1 Van Cittert-Zernike Theorem

The basic premise of astronomical imaging using radio interferometry is of correlating signals from two antennas together, along what is referred to as the baseline, as in Figure 1.1. The signal from the radio source is received by both telescopes, and the signals are then correlated together. A Fourier relationship is setup as proven by the van Cittert-Zernike theorem (Zernike 1938; Born and Wolf 1999). This states that when two points, P_1 and P_2, on a plane, are illuminated by a source at a distance R_1, and R_2, there exists a correlation between the two amplitudes and phases, despite the individual wavefronts being only partially coherent:

$$J(P_1, P_2) = \langle E_1(t) E_2^*(t) \rangle = \sum_m \langle E_{m1}(t), E_{m2}^*(t) \rangle \tag{1.1}$$

Where J represents the mutual intensity of points P_1 and P_2. E_{m1} and E_{m2} are the contributions to the wavefront at those points by divisions of a quasi-monochromatic source σ. It is assumed that contributions from different elements of the sources σ_m are statistically independent (and thus mutually incoherent):

$$\langle E_{m1}(t) E_{n2}^*(t) \rangle = 0 \tag{1.2}$$

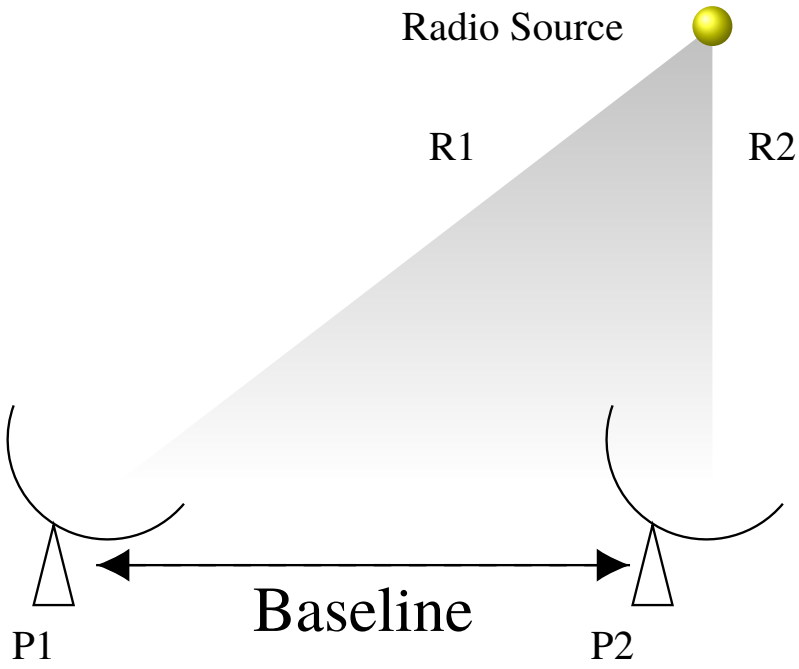

Figure 1.1: Radio Interferometer

E_{n2} is used to denote a separate contribution from the source such that $m \neq n$. This allows us to define as per Born and Wolf (1999) the "mutual intensity" function of points P_1 and P_2:

$$J(P_1, P_2) = \sum_m \langle E_{m1}(t) E^*_{m2}(t) \rangle \frac{\exp\left[ik(R_{m1} - R_{m2})\right]}{R_{m1} R_{m2}} \tag{1.3}$$

and by integrating over the whole source:

$$J(P_1, P_2) = \int_\sigma I(\Sigma) \frac{\exp\left[ik(R_1 - R_2)\right]}{R_1 R_2} d\sigma \tag{1.4}$$

where I denotes the intensity per unit area of σ. Σ denotes the centerpoint of some infinitesimal $d\sigma$ of σ. We can then derive the complex degree of coherence from this by normalisation of the Fourier transform in Equation 1.4:

$$j(P_1, P_2) = \frac{1}{\sqrt{I(P_1)}\sqrt{I(P_2)}} \int_\sigma I(\Sigma) \frac{\exp\left[ik(R_1 - R_2)\right]}{R_1 R_2} d\Sigma \tag{1.5}$$

such that $0 < j(P_1, P_2) \leq 1$ where $I(P_1)$ and $I(P_2)$ are defined as:

$$I(P_1) = J(P_1, P_1) = \int_\sigma \frac{I(\Sigma)}{R_1^2} d\Sigma = \langle E_1 E_1^* \rangle \tag{1.6}$$

$$I(P_2) = J(P_2, P_2) = \int_\sigma \frac{I(\Sigma)}{R_2^2} d\Sigma = \langle E_2 E_2^* \rangle \tag{1.7}$$

In radio interferometry this mathematical quantity represents the "autocorrelation".

With these derivations of the mutual coherence $j(P_1, P_2)$ and mutual intensity $J(P_1, P_2)$, it is now possible to relate the co-ordinate systems of the source and the measurement plane, and derive a Fourier relationship between the intensity distribution of the source and the coherence function of the two points on what is referred to as the u-v plane in radio astronomy.

The complex spatial coherence function $j(P_1, P_2)$ forms the baseline measurement unit of radio interferometry. It is now necessary to relate this optical theory to the set of baseline coordinates u, v, w from our antenna positions x, y, z relative to the coordinates on the sky l, m, n. This can be done by specifying the paths R_1 and R_2 in terms of these coordinates using routine geometry:

$$R_1^2 = (l - x_1)^2 + (m - y_1)^2 + (n - z_1)^2 + R^2 \tag{1.8}$$

$$R_2^2 = (l - x_2)^2 + (m - y_2)^2 + (n - z_2)^2 + R^2 \tag{1.9}$$

where l, m, n specify the source position on the celestial sphere, where $n = 1 - \sqrt{1 - l^2 - m^2}$, and x_1, y_1, z_1 specifying the location of P_1, with x_2, y_2, z_2 specifying the position of P_2. A straightforward expansion allows us to approximate R_1 and R_2 as:

$$R_1 \approx \frac{(l - x_1)^2 + (m - y_1)^2 + (n - z_1)^2}{2R} + R \tag{1.10}$$

$$R_2 \approx \frac{(l - x_2)^2 + (m - y_2)^2 + (n - z_2)^2}{2R} + R \tag{1.11}$$

Then the difference of these is taken:

$$R_1 - R_2 = \frac{(x_1^2 + y_1^2 + z_1^2) - (x_2^2 + y_2^2 + z_2^2)}{2R} + \frac{l(x_1 - x_2) + m(y_1 - y_2) + n(z_1 - z_2)}{R} \tag{1.12}$$

Now we can define the *visibility baseline coordinates*:

$$\begin{aligned} \bar{u} &= \frac{x_1 - x_2}{R} \\ \bar{v} &= \frac{y_1 - y_2}{R} \\ \bar{w} &= \frac{z_1 - z_2}{R} \end{aligned} \tag{1.13}$$

In most astronomical applications R is extremely large. The remaining term is a phase term that is typically disregarded in most radio astronomy texts:

$$\phi = \frac{(x_1^2 + y_1^2 + z_1^2) - (x_2^2 + y_2^2 + z_2^2)}{2R} \tag{1.14}$$

We can finally define $\bar{u}, \bar{v}, \bar{w}$ in terms of wavelengths, which is conventional in radio astronomy:

$$\begin{aligned} u &= \frac{\bar{u}}{\lambda} = \frac{\bar{u}\eta}{c} \\ v &= \frac{\bar{v}}{\lambda} = \frac{\bar{v}\eta}{c} \\ w &= \frac{\bar{w}}{\lambda} = \frac{\bar{w}\eta}{c} \end{aligned} \tag{1.15}$$

where λ represents the wavelength of the measured wavefront, η the corresponding frequency, and c the speed of light in a vacuum.

Now we can define our mutual intensity measured at a vector (u, v, w) as the radio astronomical visibility $V(u, v, w)$. By normalisation as in Born and Wolf (1999), as above, the complex degree of coherence can be similarly derived. This general optical proof is directly applicable to radio astronomy, in which the intensity distribution of our source takes the place of the sky brightness distribution, and the measurement points on a plane are our individual array antenna baselines. From this we can derive the relation between the sky and the spatial coherence measured by an interferometer; the van-Cittert Zernike theorem:

$$V(u,v,w,\eta) = \int\int I(l,m,n,\eta) \exp\left[-2\pi i(ul + vm + wn))\right] \, dl \, dm \tag{1.16}$$

$$I(l,m,n,\eta) = \int\int\int V(u,v,w,\eta) \exp\left[2\pi i(ul + vm + wn)\right] \, du \, dv \, dw \tag{1.17}$$

where $V(u, v, w, \eta)$ is the visibility distribution, and $I(l, m, n, \eta)$ is the sky brightness distribution. For the rest of our analysis we assume we are sampling a single frequency (or a frequency channel), and we can thus neglect η:

$$V(u,v,w) = \int\int I(l,m,n) \exp\left[-2\pi i(ul + vm + wn))\right] \, dl \, dm \tag{1.18}$$

$$I(l,m,n) = \int\int\int V(u,v,w) \exp\left[2\pi i(ul + vm + wn)\right] \, du \, dv \, dw \tag{1.19}$$

In the case of a flat interferometer, we can neglect the w co-ordinate:

$$V(u,v) = \iint I(l,m) \exp\left[-2\pi i(ul + vm)\right] \, dl \, dm \tag{1.20}$$

$$I(l,m) = \iint V(u,v) \exp\left[2\pi i(ul + vm)\right] \, du \, dv \tag{1.21}$$

Within interferometric imaging by aperture synthesis this is the fundamental theorem describing the relation between the sky and the visibilities at a particular frequency. By linearity of the Fourier transform, it can also be expressed using linear algebra, which can provide a more concise intuition on the results presented in Chapters 2 and 3:

$$V = FI \tag{1.22}$$

$$I = F^{-1}V \tag{1.23}$$

where F represents the Fourier transform in Equation 1.16, V represents a vector of visibilities corresponding to the vectorization of the visibility matrix $\hat{V} = E \otimes E^*$, where E represents a vector of N electric fields sampled from N antennas at some time, frequency, and polarisation; $\otimes$ represents the outer product operator and * the complex conjugation operation. The visibility matrix $\hat{V}$ is hermitian symmetric, so in practice only the upper or lower triangular matrix of $\hat{V}$ is required to compute the final sky brightness distribution $I(l, m, n)$.

1.2.2 Discrete Sampling of a Continuous Domain

The Fourier integrals defined in Equations 1.16 and 1.17 are defined on a continuous domain. However in modern radio interferometry these integrals are calculated computationally, so that digitisation onto a discrete domain is necessary. In this vein, it is first necessary to introduce the discrete analogue of the continuous Fourier transform equation defined in Equation 1.16 and 1.17:

$$I[l,m] = \sum_{i=0}^{N_{vis}} V_i[u,v,w] \exp\left[2\pi i\left[u_i l + v_i m + w_i n\right]\right] \tag{1.24}$$

This can be viewed as a many-to-many linear map, where every point in one domain maps to every point in the Fourier paired domain. The calculation of Equation 1.24 is highly computationally intractable for large N_{vis}, with a computational scaling of $O(N_{vis}^2)$. The scaling can, however, be greatly improved if the Fourier domain is sampled regularly, a procedure which can be thought of as multiplying $V(u,v,w)$ and $I(l,m,n)$ with a Dirac comb function:

$$\text{Ш}(x)_T = \sum_{k=0}^{\infty} \delta(x - kT) \tag{1.25}$$

This is also known as a Shah function; it is easily generalised to higher dimensions, designating a regularly spaced set of dirac delta functions in N Dimensions. Thus this facilitates the discretization of Equations 1.16 and 1.17:

$$V(u,v,w) = \int\int I(l,m,n)\,\text{Ш}(l,m) \exp\left[-2\pi i(ul + vm + wn))\right] dl\, dm \tag{1.26}$$

$$I(l,m,n) = \int\int\int V(u,v,w)\,\text{Ш}(u,v) \exp\left[2\pi i(ul + vm + wn)\right] du\, dv\, dw \tag{1.27}$$

This regular sampling of the two Fourier domains allows an efficient method of the Fourier transform to be used: The Fast Fourier Transform (FFT) (Cooley and Tukey 1965). The FFT has a scaling of $O(nlogn)$, making the calculation of large transforms computationally tractable where a direct implementation of the DFT would be computationally far more expensive.

Aliasing is a difficulty which arises from regularly sampling the continuous $V(u,v,w)$ and $I(l,m,n)$ domains. This is where a particular visibility measurement does not sit exactly on a grid point defined by the dirac comb function $\text{Ш}(u,v)$. The crudest method of dealing with this is to place the discrete sample of $V(u,v,w)$ at the nearest grid point. This will result in error, measured as the mismatch between the sky brightness distribution as calculated by the direct DFT, and the FFT. An example of this is shown and discussed in Chapter 3. This problem can be alleviated by the usage of an anti-aliasing function. The anti-aliasing correction can be made in grid or image space by virtue of the multiplication convolution theorem. Elizabeth

Waldram of the Cavendish Astrophysics group pioneered this approach in 1961 by using grid convolution functions in the form of gaussians, and a sinc function, however these results were unpublished. The most common approach at the present time, and as discussed in Jackson et al. (1991), is through the usage of a Prolate Spheroidal Wave Function (PSWF), which is not analytically representable (Slepian and Pollak 1961; Landau and Pollak 1961).

1.2.3 Non-Coplanar Array Imaging

Equations 1.16 and 1.17 define a Fourier integral over a three dimensional domain; this is difficult to implement in view of the regular sampling of the Fourier domains as in Equations 1.26 and 1.27. As the n co-ordinate is degenerate with the l and m coordinates by virtue of $l^2 + m^2 + n^2 = 1$, $I(l,m,n)$ can be described by a 2D domain to perform an FFT to and from.

However $V(u,v,w)$ is a true 3D domain where the w-component reflects the depth component of a particular baseline vector. Thus this breaks the "flat grid" assumption of the FFT relationship; this is not mathematically compatible with the 2D nature of the Fourier transform between $I(l,m,n)$ and $V(u,v,w)$. The w component, as discussed, is negligible in the case of a flat array observing at zenith. This combined with the projection of the celestial sphere onto a flat image means distortion would not be seen on the image.

If there is any deviation from a flat array or from observing at zenith, the result is a graduate decoherence in the image $I(l,m,n)$, especially towards higher l and m coordinates. Thus this is principally a problem in wide-field imaging where a large section of the celestial hemisphere is imaged at one time.

The major difficulties of implementing wide-field imaging due to non-coplanarity in a radio interferometer can be summarised broadly into the problems of accuracy and computational performance. For example, implementing Equation 1.24 directly would result in perfect wide-field accuracy but, due to the discussed poor scaling of this method, this method is very computationally expensive to implement for high numbers of visibilities. At the opposite extreme, executing the 2D FFT from $V(u,v,w)$ to $I(l,m,n)$ without regard for the w co-ordinate will be computationally very efficient but very inaccurate for a non-coplanar array, an array observing away from zenith, or both. All modern schemes attempt to find some compromise between computational performance and wide-field imaging accuracy. The level of accuracy required is contingent on the parameters of the observation, as well as the properties of the array such as its sensitivity and resolution.

Multiple schemes have been devised to attempt to produce accurate images whilst at the same time being efficient to implement in practice. One of the first major reviews of this topic was in Cornwell and Perley (1992), with the promotion of two ways of correcting this. One method is to extend the 2D FFT discussed above to 3D. The idea is to regularly sample the

visibilities in $V(u,v,w)$ and the sky in $I(l,m,n)$. This technique is capable of completely solving for the w-term apart from the fact that the 3D distribution of points in $I(l,m,n)$ will be mostly empty, with physically relevent solutions only existing at points where $l^2 + m^2 + n^2 = 1$. The other method put forwards is by making multiple small images of the sky, by re-phasing the interferometric data so that the image is centred on a new l_0, m_0 co-ordinate. These images, known as "facets" are then stitched together to make an overall image.

A major advance was the discovery of the w-projection algorithm (Cornwell et al. 2008). This takes advantage of the seperability of the exponential in the Fourier transform in Equation 1.17 in combination with the multiplication-convolution theorem:

$$V(u,v,w) = \int\int I(l,m,n)G(u,v,w)\exp\left[-2\pi i(ul+vm))\right]\,dl\,dm \tag{1.28}$$

where $G(u,v,w) = \exp\left[-2\pi iwn\right]$. Thus by using the multiplication convolution theorem of Fourier theory we can convolve $V(u,v,w=0)$ with the Fourier transform of $G(u,v,w)$, $\tilde{G}(u,v,w)$ to project the visibility onto a two dimensional plane, allowing efficient evaluation of the Fourier relationship through the FFT. Whilst this convolution is expensive in its own right and not particularly efficient to implement due to the inseperability of $G(u,v,w)$, it marks a great advance on previous techniques. More recent schemes take inspiration from this technique with the introduction of w-stacking as implemented in the popular imaging software *WSClean* (Offringa et al. 2014). Other schemes for correcting this problem include Image Domain Gridding (Tol et al. 2018), and optimal convolution functions as discussed in Tan (1986) and Ye et al. (2019).

1.2.4 The Dirty Beam

In a radio interferometer, the finite number of antenna elements in combination with their limited extent leads to an incomplete sampling of the spatial coherence function $V(u,v,w)$. This in turn gives rise to an image of the sky brightness distribution convolved with the point spread function or "dirty beam" of the interferometer:

$$\hat{I}(l,m,n) = I(l,m,n) * D(u,v,w) \tag{1.29}$$

This is known as the "dirty map".

An example of the incomplete sampling of the spatial coherence function is seen in Figure 1.2. To demonstrate the effect of this incomplete sampling on the final image of the sky, five simulated Gaussian sources were imaged using the DFT, with increased oversampling to emphasise the image artifacts from the dirty beam. The dirty map from this is seen in Figure 1.3, with the regularity of the VLA u,v coverage leading to particularly obvious artifacts.

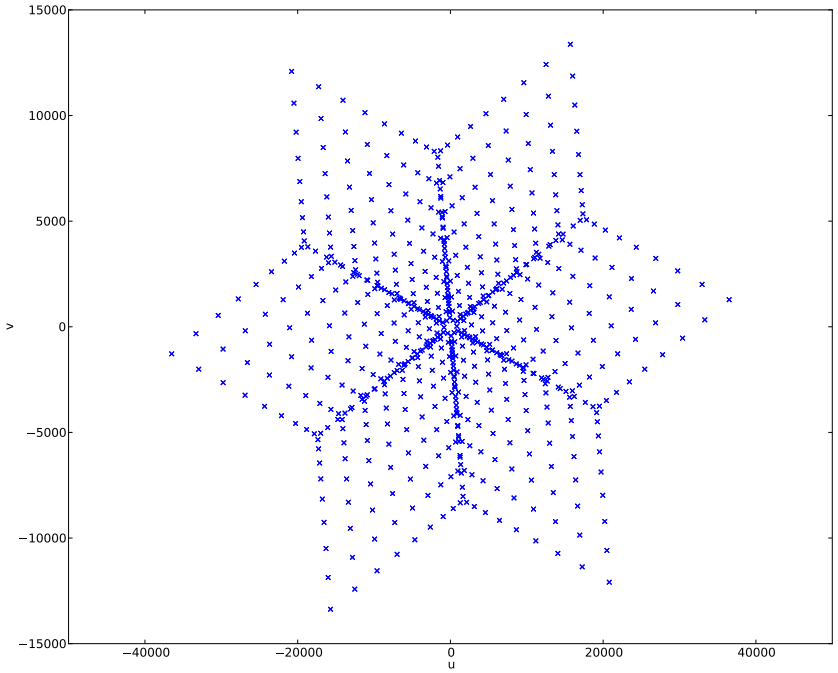

Figure 1.2: u, v Sampling of the spatial coherence ("visibility") function for the Very Large Array (VLA) in its "A" Configuration at a central frequency of 1GHz. This is a "snapshot" of the u,v,w coverage.

This dirty beam effect can be removed from the image through deconvolution, which mathematically constitutes an inverse problem. This inverse problem can be solved with various techniques such as CLEAN (Högbom 1974), and its derivatives such as Clark CLEAN (Thompson et al. 2017, pp. 555-556), and Cotton-Schwab CLEAN (Thompson et al. 2017, p. 556). Bayesian methods have also enjoyed great success, and are attractive due to their theoretical rationale, underpinned by probability theory. An example of this is the Maximum Entropy algorithm (Gull and Skilling 1984).

1.3 The Radio Interferometer

Section 1.2 gave an introduction to the mathematical theories and subtleties underlying aperture synthesis imaging with a radio interferometer. We now look into the practical details of a radio interfereometer and set out some key results regarding the scientific data products emerged.

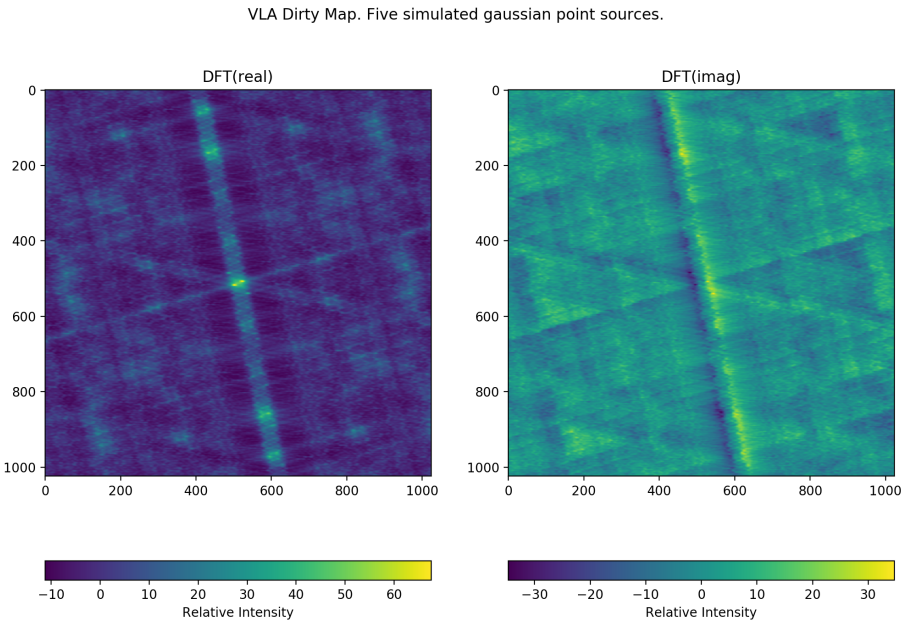

Figure 1.3: Dirty map generated for five Gaussian point sources by the DFT for the u, v coverage shown in Figure 1.2. The five Gaussian point sources are in the central portion of the map, which has been zoomed out to emphasise the effects of the point spread function from the incomplete sampling of the u, v plane.

1.3.1 Antennas and Front-End Electronics

Each receiving element of the interferometer converts an electromagnetic wave from the sky (discounting terrestial sources) to a voltage as a function of time, according to the process of electromagnetic induction described by Maxwell's Equations.

The voltage induced and its characteristics is strongly dependent on the antenna used. All antennas have a directional power pattern to them, $P(\theta, \phi)$ where θ, ϕ represent a set of spherical coordinates, such that θ is the azimuth and ϕ is the altitude of a source on the sky, the remaining spherical coordinate r can be neglected since only the far-field power pattern of the antenna is considered. This leads to a modified form of Equation 1.16:

$$V(u,v,w) = \int\int I(l,m,n) P_a(l,m) P_b(l,m) \exp\left[-2\pi i(ul + vm + wn))\right] dl\, dm \qquad (1.30)$$

sampled at a single frequency, where P_a and P_b represent the power pattern of antennas a and b. The spherical coordinates θ, ϕ can be converted to direction cosines by $l = cos(\phi)cos(\frac{\pi}{2} - \theta)$ and $m = cos(\phi)sin(\frac{\pi}{2} - \theta)$. In physical terms this results in a direction-dependent induced

voltage pattern at each antenna:

$$\mathcal{V}(t) = \int_{\sigma} E(t,\sigma)P(\sigma)\, d\sigma \tag{1.31}$$

where t represents time, and σ the solid angle in steradians on the sky. This is equivalent to a double integral over θ and ϕ.

Until now we have neglected frequency in our analysis, but in reality the power pattern is dependent on frequency:

$$\mathcal{V}(t) = \int_{\sigma}\int E(t,\eta,\sigma)P(\eta,\sigma)\, d\sigma\, d\eta \tag{1.32}$$

In addition to this, there is filtering applied to the raw voltage signals from the antenna, and the filtering varies depending on the interferometers properties such as frequency, and more practical concerns such as coupling to various noise sources around the observation site. This is often a combination of both analogue and digital filtering, with some "front-end" filtering, sometimes installed on the antennas themselves, sculpting the frequency response of the system. Cryogenic cooling is sometimes used to reduce the ambient temperature of electronics; this is especially important at higher frequencies where the temperature of the sky, T_{sky}, is much weaker than the temperature of the noise, T_{noise}. The most important filter is the band-pass filter which determines the overall frequency response of the system and is usually matched to the frequency response of the antennas in the system.

The next step is the digitisation of the continuous voltage signal to a discrete representation. Previously this was done at the central signal processing facility, but advances in low noise digital electronics in the last few years means that this step can often be done at the antennas themselves. The digitisation is often done at a low numerical precision to reduce overall data rates, with a telescope such as the Long Wavelength Array using 8-bit complex quantization (4-bit real + 4-bit imaginary), for example (Henning et al. 2010). A detailed overview and analysis of the effects of the quantization level on the sensitivity of the array can be found in (Thompson et al. 2017, pp. 316 - 347)

1.3.2 Correlation

After the digitisation of the voltages, the digitized voltages are broadcast to the central signal processing facility, known as a "correlator", where they are additionally processed. Digitisation and correlation normally take place in the same central signal processing facility. This is where a correlation operation is performed to measure the spatial coherence function defined above.

The signal from each antenna is correlated with the signals from every other antenna; with a correlation of the same antenna being known as the autocorrelation, leading to $\frac{N(N-1)}{2}$ possible

combinations. This correlation is known as the visibility:

$$V(t,\tau) = \int_{-\infty}^{\infty} \mathcal{V}_a(t)\mathcal{V}_b^*(t-\tau)\,dt \tag{1.33}$$

where $\mathcal{V}_a$ and $\mathcal{V}_b$ are the time series voltages from two antennas a and b. Through the Wiener-Khinchin Theorem, the correlation operation can be performed as a multiplication after a Fourier transform of the voltages to the frequency domain:

$$E_a(\eta) = \int_{\infty}^{\infty} \mathcal{V}_a(t)\exp\left[-2\pi i\eta t\right]\,dt \tag{1.34}$$

$$E_b(\eta) = \int_{\infty}^{\infty} \mathcal{V}_b(t)\exp\left[-2\pi i\eta t\right]\,dt \tag{1.35}$$

$$V(t,\eta) = E_a \cdot E_b^* \tag{1.36}$$

In the majority of modern correlators, the fourier transform of the voltages is the first step to be performed, where $N_{\mathcal{V}}$ digitized voltage samples from each antenna are Fourier transformed to the electric fields, forming $\frac{N_{\mathcal{V}}}{2}$ channels. Before the visibility is generated, a delay calibration step is often performed to correct for the relative mismatch in signal phase and amplitude between antennas, arising due to different lengths of cable from each antenna to the central correlator. Additional calibration solutions can be multiplied in at this point, and this is discussed in Chapter 3. After the optional calibration step the correlation in Equation 1.36 is performed for each pair of antennas. This operation can also be performed as an outer product between a vector of electric fields and its conjugate transpose, as discussed previously.

1.3.3 The Closure Phase

The visibility represents the main data product of radio interferometric imaging, allowing the sky brightness distribution to be recovered subject to the point spread function of the instrument, subsequent deconvolution, and calibration. However, Jennison (1958) noted that, by multiplying the visibilities from three baselines that form a closed geometric triangle, an additional data product, the Closure Phase, can be derived:

$$C_{abc}(t,\eta) = \arg\left\{V_a(t,\eta)V_b(t,\eta)V_c(t,\eta)\right\} \tag{1.37}$$

The major benefit of this is that C_{abc} is a true measurement of the sky without any correction for the relative antenna gains, also known as the direction-independent gains. This leads to an overall cancellation of antenna-specific direction-independent gains, such as those that arise from the antenna electronics:

$$C_{abc}(t,\eta) = \phi^s_{ab} + (\theta_a - \theta_b) + \phi^s_{bc} + (\theta_b - \theta_c) + \phi^s_{ca} + (\theta_c - \theta_a) + \phi^n_{abc} \tag{1.38}$$

Where θ_a, θ_b, θ_c are the antenna-specific phases, and ϕ^s_{ab}, ϕ^s_{bc}, ϕ^s_{ca}, are the phases from each baseline, which is a consequence of the structure of the intensity distribution on the celestial sphere. The closure phase is a true measurement of the sky, absent of any antenna based calibration terms. Thus calibration is not required to use the closure phase as an observable quantity. This mitigates any calibration errors which might occur.

Direction-dependent errors that occur, due to something such as mischaracterisation of an antenna beam, or a pointing error, are not negated through this method, and will show up as an error between the measured closure phase and true closure phase. This allows the diagnosis of errors and systematic errors in an interferometric array and this technique is used to its full power in Chapter 4.

Reconstruction of the sky from the closure phases is not directly possible, constituting an inverse problem. It is however, widely used for this purpose especially within Very Long Baseline Interferometry (VLBI), where antenna elements are seperated by such vast distances they may be considered seperate telescopes with phase calibration a major difficulty. The closure phase can contribute to calibration of visibilities and reconstruction of an image of the sky through hybrid mapping (Rogers et al. 1974). The closure phase is critical in optical interferometry due to the phase perturbations due to the atmosphere, and increases the fidelity of sky images (Baldwin et al. 1986).

1.3.4 Noise Properties

One of the most important considerations within interferometric imaging is the behaviour of the noise on individual antennas, baselines, and particularly in this thesis, the closure triads. The noise in an interferometer can be broadly split into two contributions:

- Sky Noise - Statistical fluctuations of the signals on the sky
- Thermal-Like Noise - Statistical fluctuations from the thermal noise of the antenna, electronics, and other terrestrial sources.

It is common practice to represent the units of noise in Kelvins, with each noise component having its own temperature. The sky noise will be considered as the statistical fluctuations in the antenna temperature T_a. All other noise terms resulting from the interferometer system, including receivers, electronics and the correlation operation, can be grouped under a common system temperature noise term, T_s.

A full analysis of the interferometric noise terms for different situations is presented in Appendix A, as well as providing distributions for modelling the noise. Apart from this, Appendix A also elaborates on several subtleties in the behaviour of the noise which is particularly helpful for understanding the results in Chapters 4 and 5.

1.3.4.1 Variance Properties for the Interferometer

An important step in understanding the behaviour of the noise is understanding its statistical behaviour through analysis of the variance of the noise, and resulting standard deviation. This analysis is predicated upon all noise terms being Gaussian in nature, that thermal noise between antennas is uncorrelated, and there being no cross coupling effects between antennas.

As an independent verification in the style of Kulkarni (1989), Appendix A also states the behaviour of the variance of the noise, with particular application to Chapter 5.

Here we present condensed results that are referred to throughout the following thesis. The variance of of a visibility is:

$$\begin{aligned}\sigma^2_{V_{ab}} &= \left\langle (V^S_{ab} + V^N_{ab})(V^{S*}_{ab} + V^{N*}_{ab}) \right\rangle - \left\langle V_{ab} \right\rangle \left\langle V^*_{ab} \right\rangle \\ &= \left\langle (V^S_{ab} + V^N_{ab})(V^{S*}_{ab} + V^{N*}_{ab}) \right\rangle - |V_{ab}|^2 \\ &= \left\langle |V^S_{ab}|^2 + |V^N_{ab}|^2 + 2V^N_{ab}V^{S*} \right\rangle - |V^S_{ab}|^2 \\ &= |V^N_{ab}|^2 \end{aligned} \tag{1.39}$$

where superscript ^{S} defines the contribution to the visibility V from the signal (the non-noiselike component of V), and superscript ^{N} defines the contribution from the noise. The individual antennas are defined by subscripts $_a$ and $_b$. The covariance of a baseline with another baseline sharing an antenna is:

$$\begin{aligned}\sigma^2_{V_{ab},V_{bc}} &= \left\langle (V^S_{ab} + V^N_{ab})(V^{S*}_{bc} + V^{N*}_{bc}) \right\rangle - \left\langle V_{ab} \right\rangle \left\langle V^*_{bc} \right\rangle \\ &= \left\langle (V^S_{ab} + V^N_{ab})(V^{S*}_{ab} + V^{N*}_{ab}) \right\rangle - V^S_{ab}V^{S*}_{bc} \\ &= \left\langle V^S_{ab}V^{S*}_{bc} + V^N_{ab}V^{N*}_{bc} + 2V^S_{ab}V^{N*}_{bc} \right\rangle - V^S_{ab}V^{S*}_{bc} \\ &= 0 \end{aligned} \tag{1.40}$$

The variance of a closure phase triad, C_{abc} is:

$$\sigma^2_{C_{abc}} = \frac{1}{N}\sum_{i=0}^{N}(A_i + B_i + C_i + D_i + E_i + F_i + G_i + H_i - |C_{abc}|^2) \tag{1.41}$$

$$\sigma_{C_{abc}} = \frac{1}{\sqrt{N}}\sum_{i=0}^{N}(B_i + C_i + D_i + E_i + F_i + G_i + H_i)^{\frac{1}{2}} \tag{1.42}$$

where:

$$A = V^S_{ab} V^S_{bc} V^S_{ca} V^{S*}_{ab} V^{S*}_{bc} V^{S*}_{ca} = C_{abc} C^*_{abc} = |C_{abc}|^2 \tag{1.43}$$

$$B = V^S_{ab} V^S_{bc} V^N_{ca} V^{S*}_{ab} V^{S*}_{bc} V^{N*}_{ca} = |V^S_{ab}|^2 |V^S_{bc}|^2 |V^N_{ca}|^2 \tag{1.44}$$

$$C = V^S_{ab} V^N_{bc} V^S_{ca} V^{S*}_{ab} V^{N*}_{bc} V^{S*}_{ca} = |V^S_{ab}|^2 |V^S_{ca}|^2 |V^N_{bc}|^2 \tag{1.45}$$

$$D = V^N_{ab} V^{S*}_{bc} V^{S*}_{ca} V^{N*}_{ab} V^{S*}_{bc} V^{S*}_{ca} = |V^S_{bc}|^2 |V^S_{ca}|^2 |V^N_{ab}|^2 \tag{1.46}$$

$$E = V^S_{ab} V^N_{bc} V^N_{ca} V^{S*}_{ab} V^{N*}_{bc} V^{N*}_{ca} = |V^S_{ab}|^2 |V^N_{bc}|^2 |V^N_{ca}|^2 \tag{1.47}$$

$$F = V^N_{ab} V^S_{bc} V^N_{ca} V^{N*}_{ab} V^{S*}_{bc} V^{N*}_{ca} = |V^S_{bc}|^2 |V^N_{ab}|^2 |V^N_{ca}|^2 \tag{1.48}$$

$$G = V^N_{ab} V^N_{bc} V^S_{ca} V^{N*}_{ab} V^{N*}_{bc} V^{S*}_{ca} = |V^S_{ca}|^2 |V^N_{ab}|^2 |V^N_{bc}|^2 \tag{1.49}$$

$$H = V^N_{ab} V^N_{bc} V^N_{ca} V^{N*}_{ab} V^{N*}_{bc} V^{N*}_{ca} = |V^N_{ab}|^2 |V^N_{bc}|^2 |V^N_{ca}|^2 \tag{1.50}$$

The covariance of a closure phase triad with another closure phase triad sharing an antenna is 0:

$$\sigma^2_{C_{abc}, C_{cde}} = 0 \tag{1.51}$$

The covariance of a closure phase triad with another closure phase triad sharing a baseline is:

$$\sigma^2_{C_{abc}, C_{abd}} = |V^N_{ab}|^2 |V^S_{ca}||V^S_{bc}|V^S_{de}||V^S_{bd}| \exp\left[i(\phi_{abc} - \phi_{abd})\right] \tag{1.52}$$

where ϕ_{abc} and ϕ_{abd} are the individual phases of the closure triads C_{abc} and C_{abd}.

Chapter 2

Direct Imaging: Electric Field FFT Correlator

2.1 Introduction

The work in this chapter was achieved whilst researching abroad at Arizona State University in the Summer of 2018. The correlator described here was designed, and implemented by the author, and has been published in the literature in Kent et al. (2019b), marking an interesting advance in the state of the art for imaging with radio arrays.

Current and proposed radio arrays are increasing their capability in terms of sensitivity and angular resolution through the inclusion of ever greater numbers of interferometric elements. Very dense cores of antennas provides high sensitivity at large angular scales necessary for aperture synthesis imaging of faint sources. The inclusion of longer baselines, suitably weighted versus other interferometric elements, provides resolution on smaller angular scales. The common theme however in modern astronomy is the growing complexity of arrays, and this trend is set to continue.

Future instruments such as the Square Kilometre Array (SKA; Dewdney et al. 2009b), and current instruments such as the Long Wavelength Array (LWA; Taylor et al. 2012), Canadian Hydrogen Intensity Mapping Experiment (CHIME; Chime/Frb Collaboration et al. 2018) and the Hydrogen Epoch of Reionisation Array (HERA; DeBoer et al. 2017), are looking at using high density interferometric arrays with hundreds or thousands of individual antennas to facilitate wide-field, high sensitivity and high angular resolution imaging of the sky.

There has also been a renewed focus on observations of transient phenomena such as Fast Radio Bursts (FRBs), where the origins and physical mechanisms are an active area of research. Therefore the capability to detect and image these in real-time is of key scientific importance. Interferometric measurements of FRBs have been previously achieved (Caleb et al. 2017), including by CHIME (Amiri et al. 2019b). High time resolution imaging of such phenomena would provide a significant new capability, by allowing dragnet surveys of the sky with wide field of view instruments.

Together, these two developments present a significant computational challenge for future interferometers, especially for the correlator. The standard FX correlator, where the signal from each antenna is multiplied with the signals from every other antenna to produce "visibilities", mathematically defined as an outer product, scales as $O(n_a^2)$, where n_a is the number of antennas (Romney 1985). This scaling becomes problematic as proposed arrays will contain thousands of dipole elements. All n_a^2 visibilities must be generated at the time resolution desired and subsequently gridded and then Fourier transformed to produce images, typically creating a bottleneck for high-time resolution studies.

For some array geometries, the number of visibilities calculated can be reduced by omitting short baselines with little impact on point source imaging performance. Fast convolution algorithms may also be used for correlation to further reduce the computational costs to $O(n_a^{3/2})$ (Bunton 2011).

An alternative to full-field imaging with an FX correlator is to use a beamformer that provides the telescope's response to only a few chosen locations on the sky. It does this by summing over the voltages from all antennas with appropriate delays to direct the response in a particular direction. The computational costs of a beamformer generally scale as $O(n_a)$ per calculated beam and the output data volume is proportional only to the number of beams calculated. This avoids the challenges of full-field imaging with an FX correlator, but with an associated compromise of limited sky coverage.

Direct Fourier transform imagers (Daishido et al. 1991; Foster et al. 2014) provide another alternative to both of the above approaches. Direct imaging forgoes the calculation of antenna cross products. Instead, the antenna electric fields are gridded directly onto an aperture plane and Fourier transformed into an image plane. These images can be accumulated for noise reduction, in the same way visibilities are accumulated in FX correlators.

Theoretically they can provide significant potential scaling improvement by scaling as $O(n_g \log n_g)$ where n_g is the number of grid points in the aperture, yielding a significant potential scaling advantage for high-density arrays (Morales 2011; Thyagarajan et al. 2017). Direct imagers can facilitate full-field imaging at a high time resolution because the output data volume can be much lower than for an FX correlator, scaling only as $n_g \approx n_a$ for a dense array.

Previous direct imagers such as Daishido et al. (1991); Foster et al. (2014), have relied on antennas being on a regular grid, which limits their application from a scientific standpoint. For example, their uniform layouts yield point spread functions that contain periodic grating responses that are not ideal for imaging applications. Further inherent assumptions about identical behaviour of antenna elements have to be made. As well as this, calibration still relies on using cross-correlated data products. Morales (2011) proposed the MOFF formalism as a flexible generalization of the direct imaging approach. A framework is described which exploits the computational advantages provided by direct Fourier transform imaging but with no limitations placed on the mixture of antenna elements or their placement, as well as producing fully calibrated images. In addition, provision is made for adaptive Fourier optics which can correct for non-coplanar array effects as well as antenna dependent terms. Visibilities from an FX correlator can be stored and calibrated offline due to explicit cross correlations between all antenna pairs, which is not the case for gridded electric fields. Thus, direct imagers have the added requirement to calibrate in real-time since individual antenna information is not retained after gridding. Beardsley et al. (2017) has successfully demonstrated an algorithm for this purpose.

The E-Field Parallel Imaging Correlator (EPIC), a generic implementation and simulation of this imaging approach in Python, was described by Thyagarajan (Thyagarajan et al. 2017). As a streaming, direct imaging correlator, it can be thought of as a generic, flexible real-time camera of the radio sky for large interferometer arrays.

Here, we report a GPU-accelerated implementation of EPIC, built on Bifrost, a high performance streaming framework. The implementation has been deployed on the LWA station located on the Sevilleta National Wildlife Refuge in New Mexico, USA. First light observations are shown, demonstrating its capability for transient detection. To the best of our knowledge, this is the only working example of a generic direct imaging correlator, with no restrictions on array composition or layout.

The theory of the MOFF (Modular Optimal Frequency Fourier) formalism underlying the EPIC imager is reviewed in Section 2.2, and a technical description of the implementation and development is discussed in Section 2.3. First light observations and an initial meteor transient detection are shown in Section 2.4, with benchmarks characterizing the performance on the LWA Sevilleta site discussed in Section 2.5. We summarize future work and conclude in Section 2.6.

2.2 Theory

The interferometry formulation is based on the optical theory of partially coherent quasi-monochromatic light, according to the van Cittert-Zernike theorem (Zernike 1938; Born and Wolf 1999). From this theorem a relationship can be derived between the radiation pattern on the celestial sphere (in the far field) and a spatial coherence function measured on some plane between two points sampling the radiation pattern from the celestial sphere, as in Chapter 1. This coherence function is the cornerstone of radio interferometry and is known as a 'visibility'. From Equation 1.17 there is a Fourier relationship between the sky coordinates the interferometer baseline co-ordinate system:

$$I(l,m) = \iint V(u,v,w) \exp\left[2\pi i\left(ul + vm + w\left(\sqrt{1-l^2-m^2}-1\right)\right)\right] du\, dv \tag{2.1}$$

In contrast to Equation 1.17, we deliberately neglect the w axis for now. This will be covered in more detail in Chapter 3. Mathematically, Equation 2.1 can be described in terms of an outer-product between a vector representing a single frequency channel of Fourier-transformed voltages from all antennas, and its conjugate transpose. Thus given N antennas outputting N electric field patterns in a channel, we derive a resultant $N \times N$ visibility matrix. Because of Hermitian symmetry, only the upper or lower triangle is retained for efficiency in FX correlators. This relation is as follows:

$$V_{12}(u,v,w) = E_1(x_1,y_1,z_1) \otimes E_2(x_2,y_2,z_2)^* \tag{2.2}$$

where $E(x,y,z)$ represents the electric field measured by an antenna at some location in an orthonormal co-ordinate system, with $V(u,v,w)$ representing the resultant visibility matrix. The (u,v,w) co-ordinate system represents the vector separation (baseline vector) between the different antennas.

2.2.1 MOFF

The multiplication-convolution theorem from Fourier transform theory allows us to re-arrange Equation 2.1 to give the MOFF algorithm (Morales 2011) of

$$I(l,m) = \left\langle \left| \iint E(x,y,z) \exp\left[2\pi i\left(xl + ym + z\left(\sqrt{1-l^2-m^2}-1\right)\right)\right] dx\, dy \right|^2 \right\rangle \tag{2.3}$$

where $E(x,y,z)$ is the electric field in the Fourier domain convolved with the antenna illumination pattern. It is not a point function, but an electric field distributed across some physical

extent in the measurement plane. Taking this into account Equation 2.3 becomes:

$$I(l,m) = \left\langle \left| \iint \left[W(x,y,z) * E'(x,y,z)\right] \times \right.\right.$$
$$\left.\left. \exp\left[2\pi i\left(xl + ym + z\left(\sqrt{1-l^2-m^2} - 1\right)\right)\right] dx\, dy \right|^2 \right\rangle \quad (2.4)$$

where $W(x,y,z)$ defines a 'gridding' function which constitutes a convolution in antenna space, E' represents the point measurement of the electric field within the measurement plane, and $*$ the convolution operator. In addition to the antenna illumination pattern, $W(x,y,z)$ can optionally incorporate any wide-field effects resulting from non-coplanarities in the array, as well as ionospheric effects (Morales 2011). In our implementation, we assume a co-planar array. Correcting for non-coplanarities will be discussed in detail in Chapter 3. Equation 2.4 now becomes:

$$I(l,m) = \left\langle \left| \iint \left[W(x,y) * E'(x,y)\right] \exp\left[2\pi i\left(xl + ym\right)\right] dx\, dy \right|^2 \right\rangle \quad (2.5)$$

Thus Equation 2.5 is a gridding of an electric field pattern directly for each antenna, followed by a spatial Fourier transform to produce the image. This transform is followed by squaring of the image, or cross multiplying between polarisations, and accumulating images over time. This produces what are commonly called "dirty images", which is the true sky brightness distribution convolved with the point spread function of the instrument (Taylor et al. 1999).

The EPIC architecture uses the MOFF algorithm as the basis for imaging. The computational cost of the EPIC architecture theoretically scales as $O(n_g \log n_g)$, compared to $O(n_a^2)$ for the classical FX correlator. For dense arrays, depending on array geometry, a MOFF-based correlator such as EPIC may be more efficient than an FX correlator (Thyagarajan et al. 2017). The limiting factor for the EPIC architecture is the Fourier transform size of the grid, whereas that for an FX correlator is the number of antenna pairs. A comparison of instruments and the best suited correlator type is shown in Figure 2.1 (reproduced from Thyagarajan et al. 2017).

A further characteristic of the EPIC architecture is the typically lower data rates in saving images at high time cadence (Thyagarajan et al. 2017). Images from the EPIC architecture are already calibrated and ready to enter the scientific inference process compared to visibilities from an FX architecture, which typically require additional processing offline to generate science-ready images. This means that the sky can be imaged at a higher time resolution than is possible using an FX correlator. The consequence, as will be seen, is that all-sky imaging at sub-millisecond sampling periods is feasible with EPIC, potentially yielding new insights into a wide range of time-domain phenomena at radio frequencies.

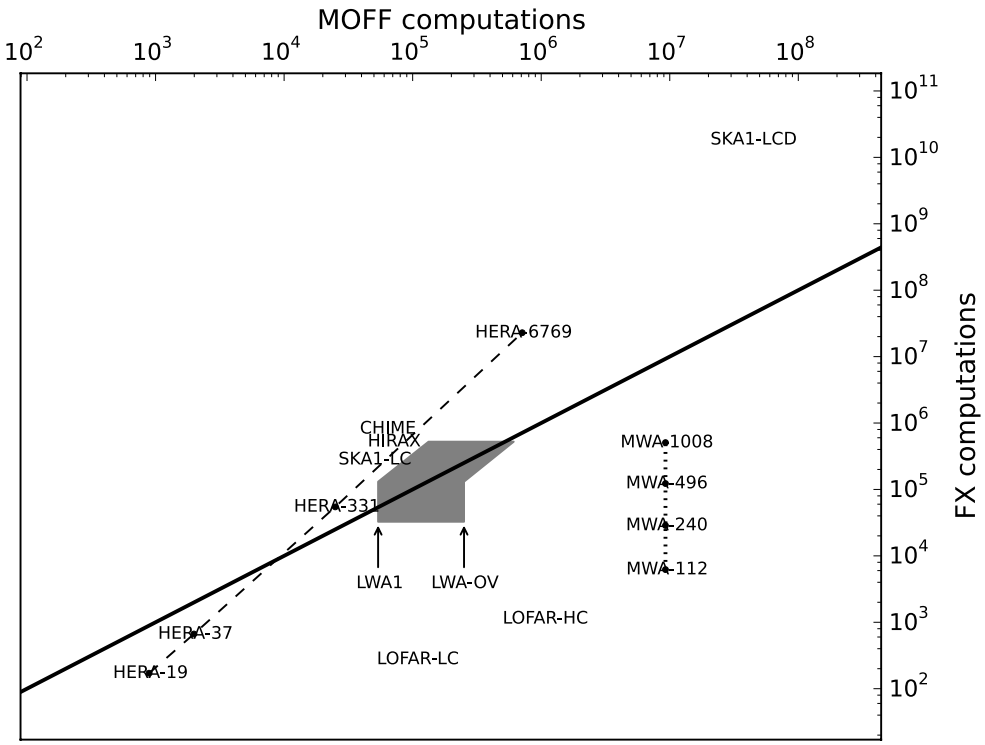

Figure 2.1: (Reproduced from Thyagarajan et al. (2017)) A selection of current and planned instruments, with the solid black line delineating the boundary in efficiency between EPIC and FX correlators. Instruments below this line are more efficient with a standard FX correlator. Above the line, EPIC is more efficient.

2.3 Implementation

2.3.1 Bifrost

The GPU-accelerated implementation of the EPIC architecture uses the Bifrost framework (Cranmer et al. 2017). Bifrost is a highly abstracted library for building high performance streaming systems. The back-end framework is built using c++ which calls high speed cuda libraries and bespoke kernels implemented by the Bifrost authors. An abstracted Python front-end is provided for ease of use.

Bifrost is based around the concept of blocks, such that each block performs some operation on the data, and then outputs it into a high-speed ring buffer in memory, which facilitates moving data between blocks. The output ring buffer from one block becomes the input ring buffer for the next block. Each block loads a "gulp" of data from the ring, and processes it before placing a gulp into the output ring. The block processes data until the input ring is emptied or the pipeline is shutdown.

Many standard signal processing techniques are implemented into the bifrost back-end with

GPU capability where appropriate. These include operations such as Finite Impulse Response (FIR) filters, Fast Fourier Transforms (FFTs), and various matrix algebra operations.

The Python front end also includes a high performance MAP function which takes a string of C++/CUDA and uses a Just-In-Time (JIT) compiler to generate and execute valid CUDA code on-the-fly using the BIFROST back-end. This provides significant flexibility in doing small mathematical operations without the need to write multiple custom blocks and implement them directly into the BIFROST framework.

Most of the EPIC implementation in BIFROST was done using the standard signal processing blocks as well as the BIFROST MAP function, with a notable exception of high-speed convolution and gridding. For this operation a custom kernel was created based on a high speed gridding algorithm developed by Romein (2012).

2.3.2 Pipeline

The real-time streaming processor implementation comprises several BIFROST blocks[a] as a Python program. All significant compute and memory operations are done seamlessly through BIFROST's high performance C++/CUDA backend. An overview of this pipeline and the relationship between the various blocks is shown in Figure 2.2.

Channelised raw data is received via UDP in a 4+4-bit complex integer format in the `UDP Receive` block. This complex integer representation serves to reduce the bandwidth required by the local ethernet connection. After the data has been captured, the channelised data is first decimated in frequency to obtain a bandwidth that can be processed without packet loss. It is then transposed to move the data ordering from $[Time, Channel, Antenna, Polarisation]$, to $[Time, Channel, Polarisation, Antenna]$. This is an important step, as it facilitates contiguous loads in the gridding convolution step.

After the `Transposition` block, the complex integer data is unpacked and promoted to a standard 64-bit complex floating-point number (32-bit real, 32-bit complex), and compensation for the signal path delays are applied using a JIT compiled BIFROST `map` function. The delay calibrated data is then convolved with the antenna illumination pattern, which is a user-defined convolution kernel. This is then gridded onto a 2D grid with a spacing of $< \frac{\lambda}{2}$, where λ is the wavelength of the channel, to ensure we are sampling all of the sky modes by sampling at the Nyquist wavelength.

This convolution and gridding operation is done using the Romein Convolution algorithm (Romein 2012) in the `Grid and FFT` block in Figure 2.2, designed specifically for high speed

[a]The source code for EPIC as well as the BIFROST-based pipeline implementation for the LWA is available alongside this thesis.

visibility gridding where locality is poor and memory bandwidth is high. The gridding convolution algorithm is described in more detail below.

Once the data has been gridded for a single time step, the gridded data is inverse Fourier transformed to produce a complex-valued image on the sky. These images are then cross-multiplied in the `Square and Accumulate Image` block to form the polarised images, which are then accumulated to a user-defined time interval depending on the science use case. After accumulation to the threshold time, the image is written to disk in a binary format and converted to a FITS image in a post-processing step. This ensures the real-time processing is not held up by high-cost image manipulation operations.

Optionally autocorrelation removal can be done to remove the zero-spacing power inherent in EPIC. Together with this, the imprint of the image of the gridding illumination kernels can be removed after the fact in a post-processing step as they are pre-generated and thus known previously.

2.3.3 Romein Convolution Algorithm

The Romein convolutional algorithm (Romein 2012) proved to be a critical step in the implementation of EPIC. Previous EPIC reference codes have attempted to use a direct convolution mapping using matrix multiplication, as described by the operator formalism in Thyagarajan et al. (2017). On a GPU, however, this results in unacceptably high memory bandwidth which causes a significant bottleneck in the code. The Romein convolution was used instead, as it is designed to reduce the GPU memory bandwidth significantly by only doing explicit memory store operations when necessary. The algorithm is designed to preferentially accumulate any grid updates into a high speed local register on the GPU core.

The Romein convolution algorithm also allows multiple convolution kernels to be combined together and applied simultaneously. This not only allows convolution of the electric field with the illumination pattern, i.e., A-projection (Morales and Matejek 2009; Bhatnagar et al. 2008), but also facilitates the inclusion of wide-field and antenna effects, such as w-Projection (Cornwell et al. 2008). Correction for non-coplanarity is discussed in Chapter 3.

The Romein convolution algorithm, written in CUDA C++, was implemented by modifying the Bifrost back-end and to add the necessary functionality. The additional Bifrost module is intended to be a generic, type-agnostic convolution kernel. This module is then called in the pipeline script from the Bifrost library using Python's `ctypes` interface.

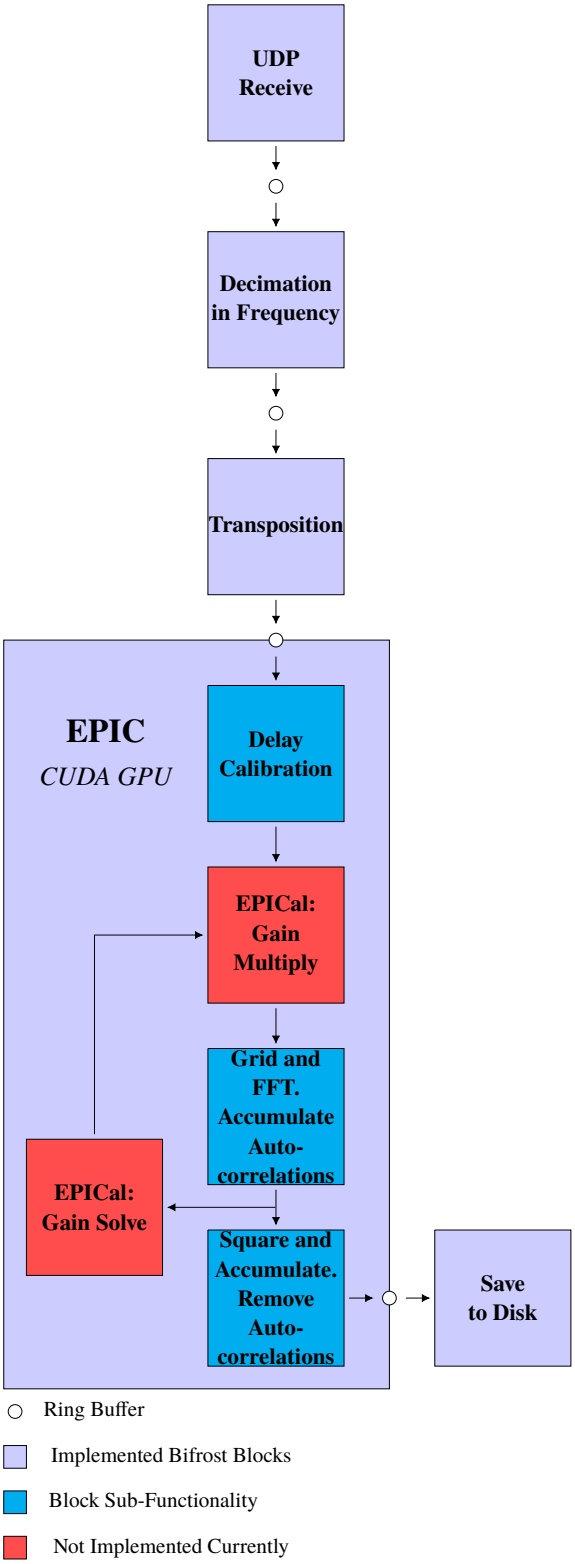

Figure 2.2: Block diagram of the BIFROST-based implementation of the EPIC architecture at LWA-SV. The blocks are named by their function and the arrows indicate the direction of data flow. The large EPIC block corresponds to a single operational block in the bifrost pipeline, with its major sub-functionality displayed. Where the calibration steps sit are also included, despite not yet being implemented. The EPIC block maps closely with the architecture discussed in Thyagarajan et al. (2017) .

2.4 Deployment and First Light

2.4.1 Long Wavelength Array

The LWA is a low frequency radio interferometer observing between frequencies of 10 to 88 MHz. It has two operational stations, one located at the Karl G. Jansky Very Large Array site and the other at the Sevilleta National Wildlife Refuge (see Figure 2.3), both in the state of New Mexico, USA (Henning et al. 2010; Taylor et al. 2012; Ellingson et al. 2013). Its high density configuration makes it an excellent candidate for deployment of EPIC. It is located at a Latitude of 34°20'54.6"N of and Longitude of 106°53'08.7"W.

The LWA Sevilleta (LWA-SV) array consists of 256 dipole antennas arranged in a dense pseudo-random arrangement inside a 110 m by 100 m elliptical aperture that is elongated north-south. An additional antenna is located approximately 300 m west of the core of the array, acting as an outrigger to help with calibration and to improve the angular resolution of the telescope. This outrigger was explicitly excluded during our implementation to ensure a high density, keeping the resultant image FFT size as small as possible.

The analog signal from each dipole is initially low pass filtered and amplified at the front end before being transmitted over coaxial cable to the electronics shelter. Inside the shelter the analog signal is further filtered and then digitized using ROACH2 boards. These use CASPER ADC16x156-8 digitiser boards to sample the dipole signals at 204.8 MHz. The digitized signals are then Fourier transformed into 4096 25 kHz channels with a time resolution of 40μs. At this point the frequency domain data, between 10MHz and 88 MHz, are requantised into 4+4-bit complex integer data, packetised, and routed over a 10/40 GbE network to a cluster of seven general purpose machines. Each machine is equipped with two Intel Xeon E5-2640 v3 processors, 64 GB of RAM, a Mellanox ConnectX-3 40 GbE network interface card, and two NVIDIA GTX 980 (Maxwell) GPUs.

2.4.2 Deployment

Initial deployment took place on the LWA-SV site during 27-31 August 2018. The EPIC architecture was deployed on a single cluster node, receiving a sixth of LWA-SV's total bandwidth. Operation of EPIC was achieved without modification to the LWA system or hardware apart from swapping the FX correlator software pipeline for EPIC. The LWA's public software library was used to perform delay calibration to account for different antenna cable lengths, and to provide the array geometry (Dowell et al. 2012). A simple square top-hat function with 3-meter extent was used as the illumination pattern for the dipole antennas. No additional calibration was performed. The observations reported here were run at an image accumulation time of 50 ms in order to allow observations of short-duration transient phenomena in the radio sky.

Figure 2.3: Aerial view of the LWA station at the Sevilleta National Wildlife Refuge. Most antenna elements are in a dense configuration towards the right of the image. A test antenna, is visible at the bottom. The signal processing hardware is contained within a modified, radio frequency shielded shipping container, visible in the left of the image.

Four channels of 25 kHz were processed with a combined bandwidth of 100 kHz centred at a frequency of 55.25 MHz. The stability of the system was tested in a 24 hour operation under the EPIC correlator. Images were generated at the raw 40 μs time cadence of the LWA-SV and then accumulated to obtain the final cadence of 50 ms. A $\lambda/2$ grid spacing was used, resulting in approximately 64^2 image pixels.

2.4.3 Detection of Meteor Transient as Proof of Concept

EPIC images the whole sky as visible to the LWA-SV station. During our initial observations, multiple small transients were identified. The majority of these are radio frequency interference (RFI), which shows up most often on the horizon, indicating a terrestrial origin. Occasionally RFI can appear overhead, reflected off of airplanes or satellites. These signals are generally narrow bandwidth and highly polarized, making them easy to recognize.

After ruling out RFI events, some physical transients were noted, the brightest of which in our observing window was a meteor striking the Earth's atmosphere, a still frame pseudo Stokes-I image of which is shown in Figure 2.4, with a light curve shown in Figure 2.5. A pseudo-stokes image is one that is formed from straightforward linear combinations of the coherency vectors from the linear polarisation parameters, but is acknowledged to not exactly represent the true

stokes vectors due to cross-coupling and polarisation leakage effects. The meteor striking the atmosphere generates a plasma, which acts as a reflector for an over the horizon analog TV transmitter at 55.25 MHz, illuminating the plasma line. This is in fact virtually identical to the methodology of studying meteor events using radar (Prentice et al. 1947). Studies of reflections such as these provide information about the speed of the neutral wind in the mesosphere and lower thermosphere through the observed Doppler shift of the reflection (Obenberger et al. 2014). The total number of meteor reflections can also be used to inform estimates of the terrestrial accretion rate (see Kortenkamp and Dermott 1998, and references therein). Such events have been observed by the LWA previously, as well as self-emission from meteor trails (Obenberger et al. 2014) and lightning (Obenberger et al. 2018). This event clearly demonstrates the potential of an EPIC system for image-based all-sky transient detection and monitoring.

2.5 Benchmarks

During the first light deployment at the LWA-SV site, performance of the present system was measured and characterised. The performance depends on both the deployment system and hardware, as well as EPIC's execution method in comparison to an FX correlator.

Overall, in the first iteration, up to 800 kHz of bandwidth is processed per GPU card on the LWA-SV correlator when running with only a single instrumental polarisation, which is useful for maximising bandwidth for faint transients, and facilitates averaging over the band. With the LWA-SV system's current hardware layout, this corresponds to 9.6 MHz of single polarisation bandwidth when EPIC is run on both GPUs of all six data capture servers. When running with both X and Y polarisations, which allows the formation of Stokes images, only half the overall bandwidth is available: up to 400 kHz per card or 4.8 MHz for the entire system. We now explore the factors contributing to the per-GPU bandwidth below and discuss ideas for improvement.

2.5.1 Maximum Throughput

To characterise the overall throughput of the system, we monitored the UDP streams being broadcasted by the ROACH2 boards running the front-end Fourier transforms and channelisation. If the system is keeping up with the input data, then there will be no packet loss. If computing requirements increase on the node, for example by increasing the number of channels per card or changing the frequency tuning such that a larger grid/FFT size is needed, then packet loss may occur as the pipeline struggles to keep up with the incoming data stream.

There are further overheads in the system, such as running a normal Linux operating system in the background that can cause occasional reductions in processing performance. To ensure

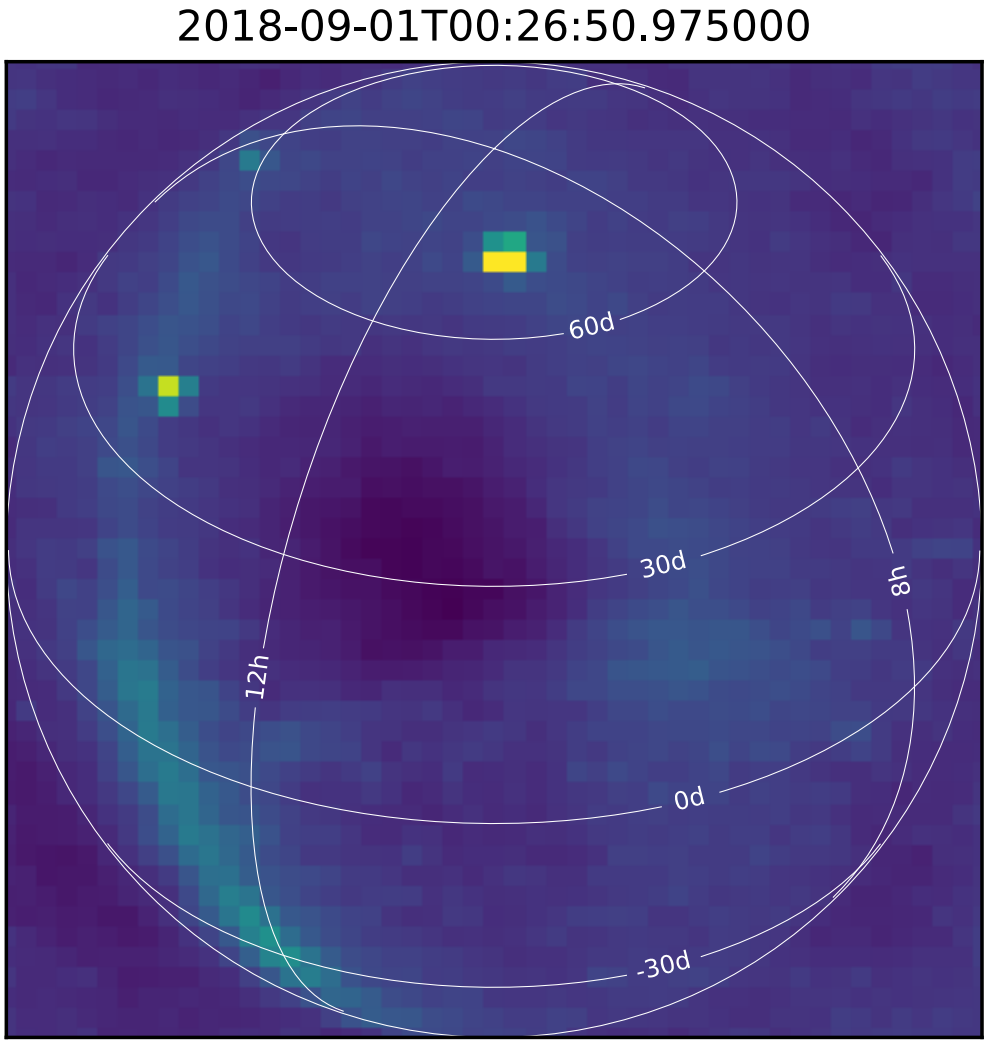

Figure 2.4: All-sky pseudo Stokes-I image showing a meteor reflection (just above the +60 declination mark) detection during an observation on the LWA-SV site (upper center). The plasma left by the meteor impacting the atmosphere reflects the signal from a 55.25 MHz TV transmitter located beyond the horizon. Lines of constant right ascension and declination in J2000 are marked in white. Cygnus A is the bright point in the upper left of the image. This observation was produced at a 50ms cadence for the images. The angular resolution of the map is 171 arcminutes, and the size of the map is 64 pixels squared.

that the pipeline does not drop packets due to such variations, we found empirically that a time 'gulp', i.e, the amount of time represented by a single chunk of data, such that the data can be processed in ≈90% of the observed time is useful, providing a 10% margin for system processing variations. For example, when ingesting 50 ms worth of data in a single gulp from the ring buffer, then in order to ensure that the system can keep up and continue running smoothly, the GPU should process it in 45 ms.

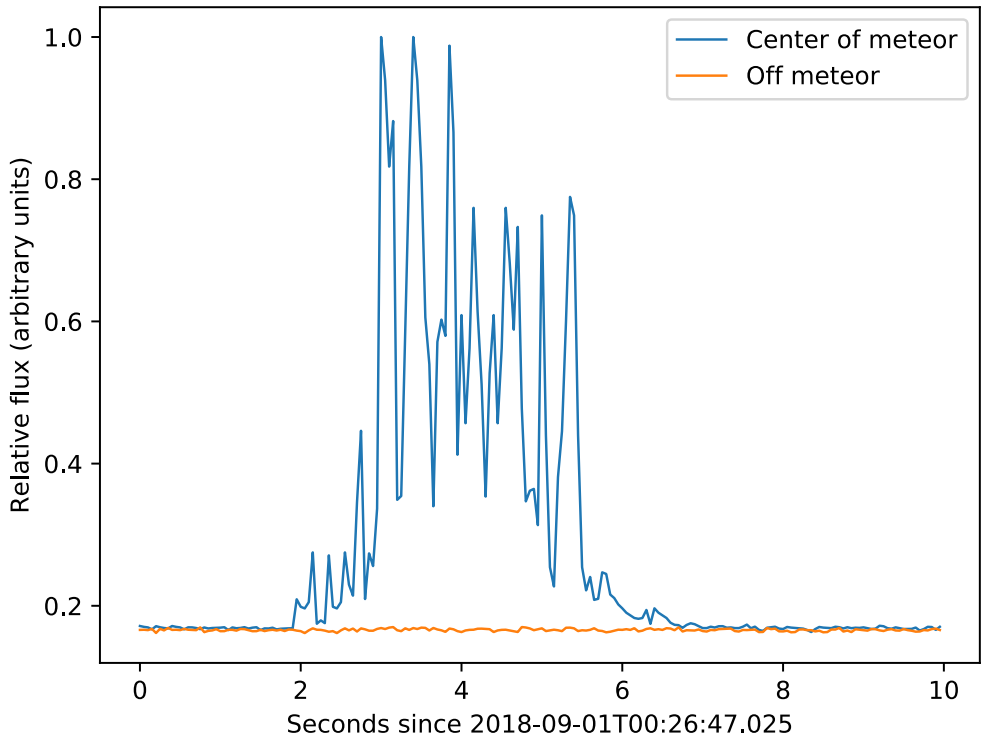

Figure 2.5: Light curve of the brightest pixel around the transient during the meteor passage noted in Figure 2.4, with a comparison to the radio background. The time resolution is 50 ms. The reflection lightcurve shows considerable structure due to changes in the plasma tail as it expands and is distorted by atmospheric winds. The lightcurve is consistent with some of the examples in Obenberger et al. (2014).

The results of our initial tests on the system are shown in Figure 2.6, where the gulp processing time and UDP packet loss fraction are shown as a function of the number of frequency channels processed. As computational resources are exceeded by increasing the number of channels, the pipeline backs up and packet loss increases, indicating that system capacity has been exceeded.

With a grid size of 64^2 and the time gulp size set to 50ms, we are capable of running up to 16 channels (400 kHz) with dual polarisations before packet loss increases to warn that the pipeline is stalling. Single polarisation mode runs more than twice as fast as the dual polarisation mode. Figure 2.7 shows the scaling of the system as a function of time gulp size when processing 100 kHz of bandwidth and dual polarisation. The scaling with gulp size is roughly linear, with the GPU coping well at a variety of representative time gulp sizes between 5 ms and 0.1 s. Similarly, Figure 2.8 shows the scaling of the pipeline with the grid size for a gulp size of 25 ms with the same bandwidth and polarization setup as in Figure 2.7. The processing times for grid sizes of 32 and 64 pixels on a side are roughly comparable, indicating that the EPIC processing time may not be dominated by the Fourier transform at these grid sizes. This can also

been seen in Table 2.1, in which the processing time per gulp is explored for a representative pipeline run. The scaling of the processing time between 64 and 128 pixels on side is around 2.5 times whereas theory would predict an increase of 4.7. The discrepancy is likely because of the underlying cuFFT library being more efficient for larger FFT sizes compared to smaller ones (Kent and Nikolic 2016). It was not possible to test longer time gulps or larger grid sizes because of insufficient memory on the GTX 980 GPUs available at LWA-SV.

2.5.2 GPU Performance

Here we assess the overall performance and suitability of EPIC for a GPU programming model. This can be explored using a roofline model, which is a common visualisation in high performance computing for analyzing the execution properties of a particular algorithm (Williams et al. 2009).

The roofline comparison between the GPU computed elements of an FX correlator and EPIC is shown in Figure 2.9. The example here is computed using a representative roofline for an NVIDIA GTX 980 GPU, used in this implementation, and for the elements of the pipeline that execute on the GPU. A GTX 2080 roofline is also provided, as a potential upgrade for the LWA correlator. The FX correlator is clearly in the compute bound regime, whereas EPIC is memory bound. This means increasing the memory bandwidth available rather than compute power will be more beneficial for EPIC, in contrast to a FX correlator which is predominantly compute-bound.

Upgrading of the LWA-Sevilleta correlator to use GTX 2080's, which have over double the memory bandwidth and compute performance, should lead to a significant increase in the bandwidth that can be processed. Assuming an increase in performance of at least two times with the new cards, not withstanding additional optimisation, it should be possible to process more than 15 MHz of LWA bandwidth.

The direct comparison is not wholly appropriate, however, for two reasons. First, the EPIC architecture provides end-to-end real-time imaging (from raw antenna voltages to science-ready calibrated images), whereas an FX correlator predominantly consists of a single mathematical operation, namely, the outer product of the raw voltages and therefore does not implement calibration or imaging, which incur additional costs. Second, if fast time-domain studies (timescales $\lesssim$ 1 ms) are to be performed with an FX-based correlator, the cost of gridding and imaging will be much higher since they have to be performed at such fast cadences and have not been included in the present estimates.

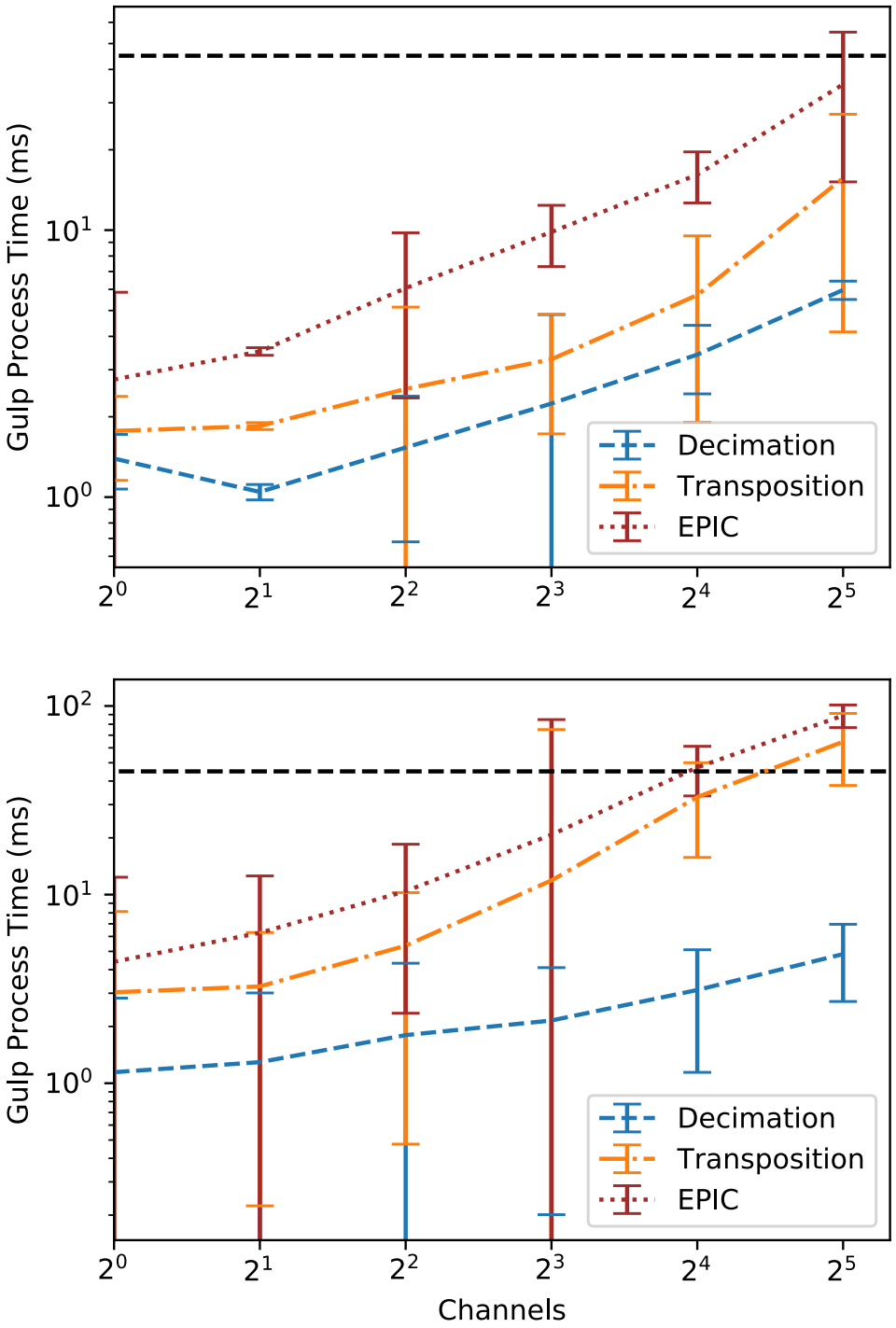

Figure 2.6: Processing time as a function of the number of channels being processed for (top) single polarisation and (bottom) dual polarisations, with EPIC running on LWA-SV with a time gulp of 50 ms and grid size of 32 pixels on a side. At 32 channels is when we began experiencing packet loss on dual polarisations, on the incoming data stream carrying electric field data, which marks when the system is no longer able to keep up with the input data rate. The black dashed horizontal line denotes the 90% processing time for the gulp size.

Block	Processing Time (% of gulp time)
Decimation	2
Transposition	4
EPIC	90
EPIC – FFT	35
EPIC – Gridding	20
EPIC – Cross-Multiply	40
EPIC – Data Transformation	5
Save	4

Table 2.1: Representative approximate breakdown of processing time by block as a fraction of the time gulp. This was done for a grid size of 64^2, 2048 40μs time samples and eight channels.

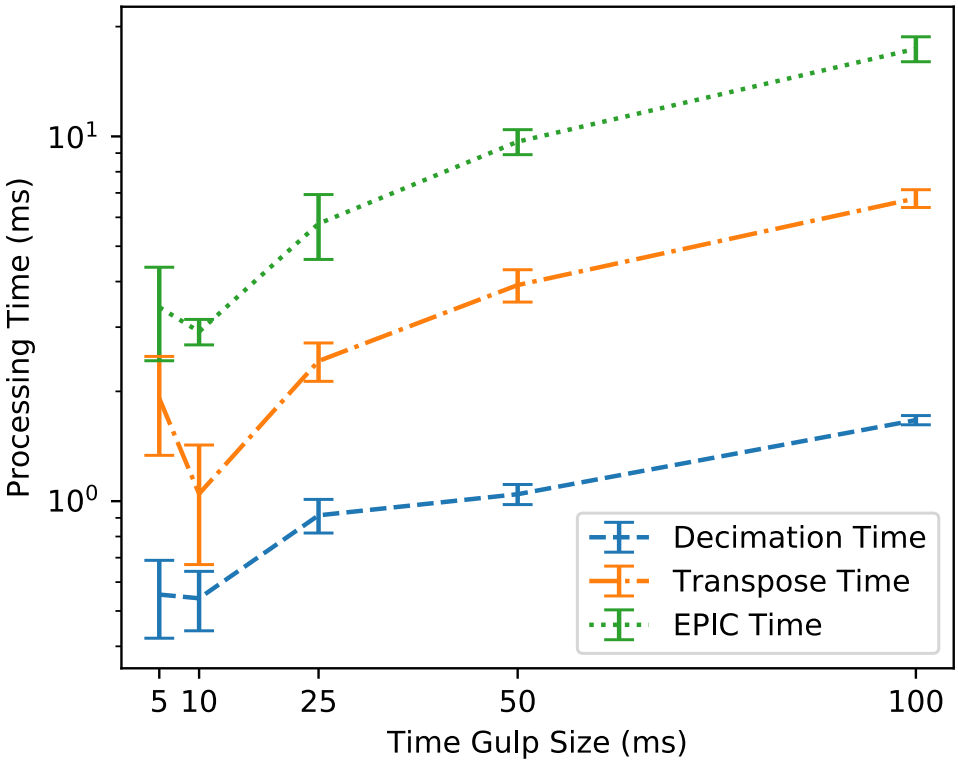

Figure 2.7: An exploration of how the system scales as a function of the time gulp sizes for 100 kHz of bandwidth and dual polarization. Each vertical bar is sub-divided to show the time used by each block in the pipeline. The legend corresponds to the blocks in Figure 2.2, with 'Image and Accumulation' corresponding to the pipeline element on the CUDA GPU. This data was derived from at least 600 trials of each time gulp size.

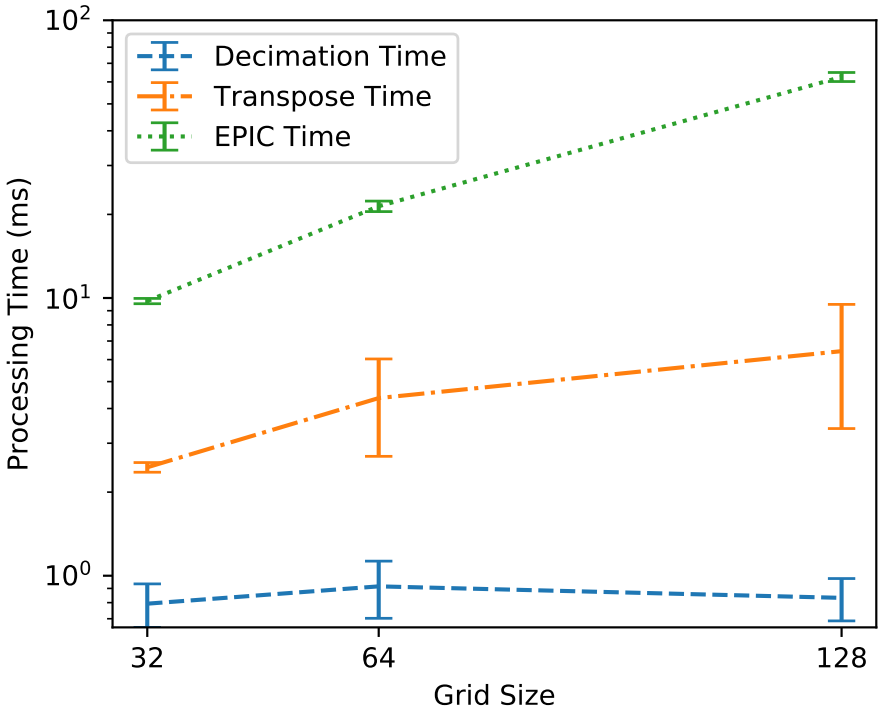

Figure 2.8: An exploration of how the system scales as a function of the grid size for a time gulp of 25 ms, 100 kHz of bandwidth, and dual polarization. The Grid Size is the size of one dimension of our squared grid. This data was derived from at least 600 trials of each grid size. 'EPIC Time' in this instance is the GPU element specified in Figure 2.2. A grid size of 128 is more than can be processed in real-time by the system, as it causes packet loss on the input UDP stream, but it is plotted here to show scaling.

2.6 Discussion

The first version of a working EPIC direct imager, through direct implementation of the MOFF algorithm, has been developed in full, and its implementation and operation demonstrated at the LWA-SV site. Observations of transients from reflections of terrestrial transmissions off passing meteors on timescales of $\sim$ 2 s at a cadence of 50 ms are reported. These successfully verify EPIC as a science-capable interferometric imaging capability. Whilst the LWA-SV is limited in terms of its sensitivity and resolution, especially in the absence of a full calibration system, it served as an excellent platform for deployment of this technology due to its extraordinary phase and gain stability, as well as a flexible correlator back-end.

The Bifrost framework aided in implementing EPIC. The c++/cuda back-end abstracts away complicated constructs, such as the ring buffers, which form the communication backbone between processing steps. The major advantage is the native cuda support, facilitating access

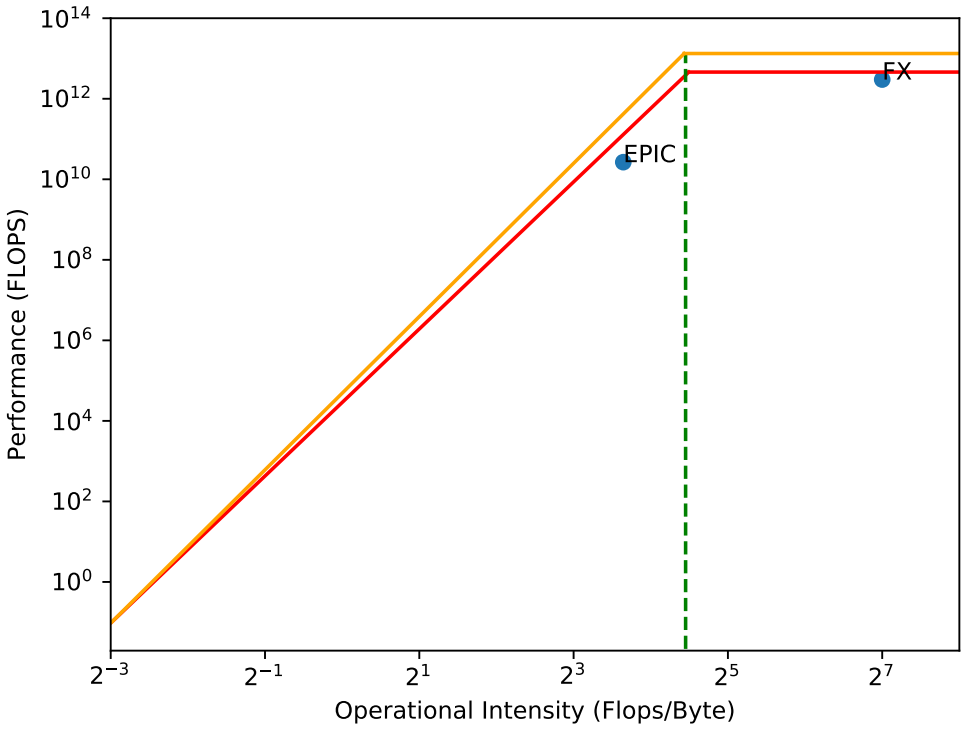

Figure 2.9: A roofline model comparing an FX correlator outer product with the EPIC pipeline on a NVIDIA GTX 980 GPU. The red line represents the maximum number of operations per second for a GTX 980 GPU, and the orange line for a GTX 2080, a potential upgrade card for the LWA. The rooflines for both cards were worked out using the memory bandwidth and peak compute performance for single-precision floating point (FP32) operations. The memory bandwidth of the GTX 980 and GTX 2080 are 226 GB/s and 616 GB/s, with peak FP32 performance of 4.6 and 13.3 TeraFlops, respectively.

to the power of the GPGPU paradigm. Extending the Bifrost framework, such as adding extra GPU-enabled processing blocks, was straightforward.

This successful deployment and working demonstration of the principles behind EPIC and the MOFF formalism mark a paradigm shift in correlator technology. The impact is particularly great for high density arrays such as SKA1-Low and the completed HERA configuration. In further instrumental development this work offers also the capability to provide a self-triggering transient survey instrument in arrays such as the Low Band Observatory (ngLOBO; Taylor et al. 2017) and the LWA Swarm Telescope (Dowell and Taylor 2018). Higher frequency instruments, such as the MWA, can benefit from the unique capabilities of EPIC for exploring FRB phenomena, given its ability to image the entire celestial hemisphere simultaneously at high time cadence. Further possible scientific uses range from ionospheric disturbance mapping

to direct observations of compact astrophysical sources.

Although this work has successfully demonstrated direct imaging, by generating real-time images through the use of the antenna electric field measurements, several significant hurdles remain en route to a fully capable implementation. The greatest hurdle is accurate and dependable calibration, which must necessarily be done in real-time due to the nature of this form of interferometric imaging. With the LWA-SV telescope, it was possible to take advantage of the high degree of phase stability to form coherent aperture synthesis images with only correction for the cable delay gains. However, as pointed out in Beardsley et al. (2017), the dynamic range of the instrument can be greatly increased with the implementation of a scheme to solve for the relative antenna gains and phases so as to correct for the time varying element of these. An extremely important part of the implementation of EPICal is real-time detection and excision or flagging, of radio frequency interference, which could severely impact calibration solutions if not correctly taken into account.

The flat grid assumption has been used in this work, neglecting the w term in the interferometry measurement equation. This is technically an incorrect implementation of the equation and the matter is considered in detail in Chapter 3, in which various types of w-correction are discussed. It is concluded that there is a false equivalence between correcting wide-field effects in visibility and in direct electric-field imaging.

There is also a need for highly optimised software implementations. The direct imaging correlator described here has undergone rudimentary optimisations, but significant work remains to be done in optimising computing time usage, such as the FFT step, the Romein gridding step, and the generic map kernels. In particular, the map kernels take up a large amount of computing time due to their highly generic nature, which precludes any significant optimisation outside of hand writing and optimising CUDA C++ kernels, a process which is highly time-intensive. Some progress has nevertheless been made in this direction, with the major benefit being an increase in the amount of bandwidth that a computational node within a correlator can process. This brings corresponding benefits for the dynamic range, and situations in which large amounts of frequency bandwidth are required.

Finally, this technique has been shown to provide an extraordinarily powerful technique for radio transient analysis, with its ability to image the whole sky at high time resolution. Some transient phenomena, such as bursts, are fleeting and rare, and it is therefore important to be able to detect these bursts automatically, with consideration given to reducing the false positive rate from radio frequency interference, and correlation with other imaging modes.

Inclusion of these features will further increase EPIC's scientific repertoire, through correction of antenna based terms in the imaging process, as well as precision calibration in real-time, yielding precision astronomical observations across the whole sky, with high time resolution.

These innovations are expected to be implemented in the next iteration of development of EPIC on the LWA. Observations using this direct imaging mode will provide the capability for source-agnostic blind transient surveys, limited only by the sensitivity and dynamic range of the instrument, as well as deep integration for cosmological observations such as the Epoch of Reionisation.

Chapter 3

Direct Imaging: Wide-Field Imaging

3.1 Introduction

In Chapter 2, the deployment was described of a functioning implementation of the real-time E-Field Parallel Imaging Correlator (EPIC) (Thyagarajan et al. 2017). This is a new capability in radio interferometry using direct electric field imaging (Kent et al. 2019b). It allows interferometric images of the sky to be made directly from the electric fields of each antenna, without forming visibilities and without restriction on antenna placement or the homogeneity of array elements. This reduces computational cost as well as enabling high time resolution wide-field surveys of the sky. This demonstration was performed on the Long Wavelength Array (LWA) (Taylor et al. 2012), using a direct fourier imaging mathematical formalism (Morales 2011) in the form of EPIC. The work in this chapter has been published in Kent et al. (2019a).

The major benefit of using the MOFF formalism over the traditional FX correlator (Thompson et al. 2017, pp. 353-375), is that MOFF is able to reduce the computational scaling from $O(n_a^2)$ to $O(n_g \log n_g)$, where n_a is the number of antennas, and n_g is the number of grid points in the aperture. This is because direct imaging does not involve a costly outer product operation where the visibilities are correlated, with $O(n_a^2)$ scaling, between the vector of channelised electric fields and its transpose. Direct imaging therefore provides significant scaling benefits for dense arrays with large numbers of antenna elements.

This provides the capability for wide field of view imaging in real time at high time resolution, for the first time in time domain radio astronomy. This technique can be deployed for the investigation of Fast Radio Bursts (FRBs) on interferometers (Caleb et al. 2017), which

have been recorded at low frequencies using instruments such as the Canadian Hydrogen Intensity Mapping Experiment (CHIME; Amiri et al. 2019b).

In the original demonstration of the functioning EPIC correlator on the Sevilleta station of the LWA, non-coplanarity was not corrected for. Images were formed by gridding the electric fields in a convolution step, and then Fourier transforming using the Fast Fourier Transform (FFT) algorithm (Thyagarajan et al. 2017). The Sevilleta LWA site is a fairly co-planar array, and at the low frequencies that the station operates at this does not lead to great distortion of the image; this is apparent in the commissioning images in Chapter 2 and Kent et al. (2019b), demonstrating the high time resolution capacity.

In addition to the problem of non-coplanarity, the approach in Chapter 2 has its limitations in that its applicability is restricted to very dense arrays where the antennas fit within a finite sized grid so that the FFT algorithm can be utilised. The usage of the FFT also restricts the resolution and placement of pixels on the sky. Thus whilst the direct imaging formulation does work very well as described in Chapter 2, these are undesirable characteristics for scientific cases where flexibility is prized.

For accurate wide field imaging, it is necessary to correct for wide field effects arising from a non-coplanar measurement plane (Cornwell and Perley 1992). With EPIC, the non-coplanarities must be solved and corrected for in real-time. Various approaches have been used to solve for the w-term in the van-Cittert Zernike equation for visibility based imaging, such as w-Projection (Cornwell et al. 2008), and w-stacking (Offringa et al. 2014). Use of a traditional w-stacking method for solving for the non-coplanarities will be demonstrated and a cost analysis performed. Unfortunately, although the mathematics will be shown to be identical, correction for wide-field effects using electric fields is difficult to implement in real time.

Finally, direct radio imaging using a Direct Fourier Transform (DFT) matrix with the digitized electric field data is shown to be a computationally tractable solution to the problem of real-time high time resolution wide field imaging with some additional attractive properties. The DFT approach places no restrictions on the location of the antennas, thus allowing direct imaging of sparse arrays. It additionally gives fine grained control over the pixel locations on the sky. This overcomes some of the major limitations of the original approach demonstrated in Chapter 2.

The DFT approach discussed here is reminiscent of beamforming techniques, where the antennas in the array are coherently summed into a beam to maximize the gain in a particular direction on the sky. This technique allows multiple phase centers for later correlation. It can also be used for reduction of data volume compared to correlating all antenna elements, by additively beamforming between a set of antennas and then correlating voltage beams from each beamformed set of antennas. This is known as Phased Array beamforming and can be

used to reduce data rates and carefully sculpt beams in an interferometer, which is planned for the SKA-Low interferometer (Adami and Turner 2011). Beamforming can be done in both the voltage and frequency space, depending on the technique (Barott et al. 2011). Beamforming can be done as a dedicated observation mode for an antenna, for example in pulsar observations as at LOFAR (Mol and Romein 2011). It can also be used in concert with correlative techniques as described above. A related technique is the idea of tied array beamforming, as implemented on the MWA (Ord et al. 2019) and LOFAR (Mol and Romein 2011), where voltages are loaded from disc and beamformed to a few pixels on the sky, creating a set of steered voltage beams.

Beamforming can also be done using an FFT algorithm (with $O(n_g \log n_g)$ scaling) with a redundant layout of antennas to facilitate fast beamforming (Masui et al. 2019). This allows multiple antenna beams to be formed with different pointing angles, which can be monitored in real-time for transient detection, such as at CHIME (Amiri et al. 2019b). Examples of arrays that use a combination of these beamforming approaches include the LWA (Taylor et al. 2012), MWA (Tingay et al. 2013; Ord et al. 2019), and CHIME (Bandura et al. 2014).

The technique described here however is a direct implementation of the interferometry measurement equation, satisfying the same mathematical relationship as imaging with visibilities. It can be interpreted as a convergence of beamforming and imaging; the two approaches are equivalent to each other with careful calibration. An imaging approach based on the FFT, discussed in Chapter 1, can be seen as analagous to beamforming to all pixels in the selected field of view simultaenously, through usage of the FFT algorithm. The DFT approach can also be interpreted in this way, where we have substituted the FFT (of visibilities or electric fields) for a DFT. The DFT formalism facilitates a direct imaging correlator which can operate in real-time at high time resolution in a similar vein to EPIC, but with the substitution of a grid and FFT step for a multiplication with a DFT matrix.

Use of a DFT matrix means that a sparsely distributed array can be used for direct electric field imaging, as it is released from the grid size constraint of the FFT algorithm to maintain real-time performance, such as in the EPIC correlator. Additionally any sky coordinates can be sampled at any resolution, thus allowing high pixel resolution images of the sky with a selectable field of view to be generated. True angular resolution is still limited by the dirty beam of the interferometer.

The DFT matrix can also incorporate time dependence, allowing exact tracking of sources across the sky with precise pixel placement with imaging at very high time resolution as in Chapter 2. This has consequences for time domain astronomy looking at radio transients. For example, using this approach, a known FRB repeater such as those described in Amiri et al. (2019a) could be tracked with multiple pixels to understand both the spatial and temporal properties of the source; this technique is more capable than voltage beamforming to sparse

locations on the sky, which in practice only allows temporal and broad spatial information to be recovered. Additionally this approach has great potential in solar observations, jovian observations and exo-planetary observations.

An overview of the theory dictating the MOFF formalism and wide field correction with electric field based imaging, including the DFT, is shown in Section 3.2, with an analysis of using a w-stacking and w-projection technique shown in Section 3.3 and why it is ultimately difficult to implement in a practical direct imaging telescope. Use of a Direct Fourier Transform in real-time with data from the LWA at Sevilleta (LWA-SV) is shown in Section 3.4, along with a detailed performance analysis showing why the surprising high performance of the direct Fourier transform method is possible.

3.2 Theory

Direct radio imaging, such as EPIC (Thyagarajan et al. 2017), takes advantage of the multiplication convolution theorem to re-arrange the canonical van-Cittert Zernike theorem into a Fourier relationship between the electric fields and the sky brightness distribution:

$$I(l,m) = \left\langle \left| \iint E(x,y,z) \times \exp\left[2\pi i\left(xl + ym + z\left(\sqrt{1-l^2-m^2}-1\right)\right)\right] dx\, dy \right|^2 \right\rangle. \tag{3.1}$$

where the electric fields are measured at antenna locations in u, v, w, which are the physical locations of the antennas in the local coordinate frame of reference in units of wavelengths at the sampled frequency. The electric fields are convolved with the antenna illumination pattern onto a grid at this location, a Fourier transform is performed, and the resulting matrix is squared with its complex conjugate, and accumulated over N_T timestamps. This is exactly the same as a "dirty" image formed with visibilities.

The above equation has the same non-coplanarity w-term (Cornwell and Perley 1992) as exists with visibility based imaging. This intuitively makes sense as the visibilities are the cross-correlations of the electric fields, represented as an outer product of the vector of electric fields and their complex conjugates:

$$V_{12}(u,v,w) = \left\langle \overrightarrow{E_1}(x_1,y_1,z_1) \otimes \overrightarrow{E_2}(x_2,y_2,z_2)^* \right\rangle \tag{3.2}$$

The electric field contributions from the sky can be modelled as

$$E(r) = \int E(\Omega_k) e^{\phi} e^{-i\vec{k}\cdot\vec{r}}\, d\Omega_k \tag{3.3}$$

and visibilities as:

$$V(r) = \int I(\Omega_k) e^{-i\vec{k}\cdot\vec{r}}\, d\Omega_k, \tag{3.4}$$

where $\vec{k}$ is a vector of sky cosine coordinates and $\vec{r}$ a vector of measurement plane coordinates. Here, $E(\Omega_k)$ represents the electric and $I(\Omega_k)$ the intensity pattern respectively at a particular location on the sky. We integrate over the solid angle elements $d\Omega_k$; e^{ϕ} is a random phase term indicating that all points on the complex sky are largely incoherent with respect to each other.

These equations both satisfy the Helmholtz equation, and therefore comprise a valid wave equation. This property, along with Equation 3.1, suggests that the same approaches to wide field correction should apply for electric fields, as the same classes of solution are allowed. Thus, any method valid for w-correction with visibilities might also work with electric fields since they use the same set of basis functions. Consequently, various techniques are available to correct for non-coplanarity, such as w-Projection (Cornwell et al. 2008), w-stacking (Offringa et al. 2014) or optimal gridding functions (Ye et al. 2019). It may also be possible to correct for the non-coplanarity by using an expansion with spherical harmonics and this is discussed in Appendix B.

3.2.1 Direct Fourier Transform Operator

Use of the direct Fourier transform is by far the easiest method, but it suffers from poor scaling as the image size increases. Fortunately, this is significantly improved for electric fields rather than visibilities. With electric fields, the scaling is $O(N_K N_A)$, compared to $O(N_K N_A^2)$ with visibilities; here, N_K is the number of sky pixels (l,m coordinates), and N_A is the number of antennas.

The Fourier relationship from the electric fields can be viewed as a bilinear map from the electric fields to the dirty map space, where the dirty map is the true sky convolved with the dirty beam of the instrument:

$$S = \left|DX\right|^2, \tag{3.5}$$

and S is defined as the real matrix representing the sky-modes sampled at a discrete set of sky cosine coordinates. Here, D is the complex DFT matrix representing the direct Fourier transform of the matrix X which is our electric field data matrix. The absolute value squared in Equation 3.5 indicates that the square of the absolute magnitude of each complex entry in the DX matrix has been taken.. This is equivalent to a Hadamard product between DX and its complex conjugate $(DX)^*$. The DFT matrix D is of the form:

$$D = \frac{1}{N}\begin{bmatrix} \omega_{0,0} & \omega_{0,j} & \dots & \omega_{0,N_A} \\ \omega_{i,0} & \ddots & & \omega_{i,N_A} \\ \vdots & & \ddots & \vdots \\ \omega_{N_K,0} & \dots & & \omega_{N_K,N_A} \end{bmatrix}, \tag{3.6}$$

where $w_{i,j} = exp\left[-i2\pi k_i \cdot r_j\right]$, with i representing the index of the sky cosine coordinate being sampled, and j the index of the antenna from which the electric fields are being sampled.

The matrix S has dimensions of N_K rows and N_T columns, where N_T is the number of timestamps. The D matrix can be considered time-independent for short observations and can therefore be pre-computed, however time dependence is easily incorporated. Its dimensions are N_K rows and N_A columns. The X matrix has dimensions of N_A rows and N_T columns. This allows batch imaging of multiple timestamps.

One of the bonuses of using the DFT over an FFT is that there is flexibility in which sky pixels are sampled due to not being held to the requirement of a regular grid. It also places no limitations on the placement of the antennas in the measurement plane, whereas with the FFT the limitation is the finite grid size, and computational cost rises as this grid size increases.

This property of the DFT allows sparse arrays to be imaged using the EPIC correlator. By re-generating the D matrix during an observation, thereby explicitly adding time dependence, different observation modes can be incorporated such as tracking celestial objects and imaging them at high time cadence and high resolution.

Within the original EPIC architecture described by Thyagarajan et al. (2017), it is shown that EPIC is a generic framework that allows for optimal image making with heterogenous arrays. This is where the antenna's have different properties such as:

1) Cable Complex Gains
2) Antenna Complex Gains
3) Antenna Illumination Pattern

These are still able to be to dealt with using operator D in Equation 3.5, which can have these terms incorporated. For example take 1). The complex gains from the cables should be known from the characterisation of the instrument, and are direction independent. Thus the correct gains and phases can be applied to the D matrix. This can be modelled as a Hadamard product between a matrix A_c and D:

$$\bar{D} = D \odot A_c. \tag{3.7}$$

where the rows of A_c are identical in each column, but the columns differ, corresponding to the gains and phases of each individual antenna. For 2) this is also a position independent term but naturally has a time dependence associated with it and must be solved through calibration of the system. An example of calibrating an electric field based direct imager has been demonstrated by Beardsley et al. (2017), and this can be folded into a separate complex A_s matrix, with the caveat that there is now a time dependence as the calibration solutions naturally change over time, and this is also multiplied point-wise with the above:

$$\bar{D} = D \odot A_c \odot A_s. \tag{3.8}$$

It is more difficult to take into account 3), as it is position dependent and antenna dependent. Thus there will be another matrix defined, A_i representing the electric field patterns, the Fourier transform of the antenna illumination pattern, at each sky cosine coordinate and antenna. Again we can do another point-wise multiplication:

$$\bar{D} = D \odot A_c \odot A_s \odot A_i. \tag{3.9}$$

Thus Equation 3.5 becomes:

$$S = \left|\left(D \odot A_c \odot A_s \odot A_i\right) X\right|^2. \tag{3.10}$$

The antenna beam correction is done in the sky space compared to in the measurement plane; correction in the measurement plane is the method described in Thyagarajan et al. (2017). This is mathematically equivalent to techniques such as Aw-Projection (Bhatnagar et al. 2008).

The DFT approach is therefore equivalent to the MOFF formalism shown in Morales (2011), with several key differences. The major difference is that the scaling is no longer $O(n_g \log n_g)$ because we are no longer convolving onto a regular grid and using the FFT, which is the architecture used previously (Thyagarajan et al. 2017). It is now $O(N_K N_T N_A)$, so that for arrays with many antennas or which produce images with many different sky positions sampled the cost increases linearly for each dimension. This will be demonstrated to be computationally fast enough to run a real-time direct radio imager using data from LWA-SV.

The many benefits of this approach include that there is no longer any constraint to use a small dense array, as required with the original EPIC formulation. The antennas can be located anywhere (with additional consideration required for ionospheric behaviour), and the sky can be sampled at any location. Thus it is feasible to make high time resolution images of the sky using the electric fields directly by taking advantage of the speed of the matrix multiplication in Equation 3.5 implemented on modern GPU hardware.

3.3 *w*-Projection and *w*-Stacking

Using methods similar to visibilities for correction of the w-term were considered, and are tested in detail here to explore whether they are both mathematically capable of correcting for the w-term with electric fields, as well as if they are efficient to implement in practice. Specifically, w-projection (Cornwell et al. 2008) and w-stacking (Offringa et al. 2014), were explored.

In the original formulation of EPIC, a convolution was performed which mapped the electric fields to the measurement plane, and then an inverse Fourier transform was applied using the FFT algorithm.

To apply w-correction in this method, we could convolve the electric fields with a Fourier transformed Fresnel pattern, the w-kernel, in the same way that w-Projection works (Cornwell et al. 2008). Unfortunately the grid sizes in EPIC are often small to account for dense arrays and to constrain the computational cost of the FFT. This means that the size of convolutional kernel is limited by the grid size. The size of the w-kernel can be determined as in (Mitchell and Bernardi 2014):

$$\text{Kernel Size(1-D)} = \left[\left[\frac{w_{max}\theta}{2}\right]^2 + \left[\frac{w_{max}^{\frac{3}{2}}\theta}{2\pi\epsilon}\right]\right]^{\frac{1}{2}}, \tag{3.11}$$

where θ represents the field of view size, w_{max} the maximum w value, and ϵ the fraction of peak to represent the w pattern out to. With a field of view set to the entire sky, ϵ set to 0.01, and a w_{max} set to 2.66, similar to no-coplanarity in LWA-SV, this results in a recommended convolution size of more than 10, which is impractical for small grid sizes used at low frequencies, such as those used in the implementation of the EPIC correlator at LWA-SV (Kent et al. 2019b). However at higher frequencies where grid sizes increase in the EPIC correlator (see Thyagarajan et al. 2017) this is a perfectly acceptable convolution size.

An alternative is to use a similar approach to that in WSClean (Offringa et al. 2014) using w-Stacking, where we grid each electric field at a particular “plane", where the planes are spaced out in w, creating a stack. The process for correction in this way is:

1) Sort electric fields in order of increasing/decreasing w.
2) Apply electric field to nearest w-plane. Optionally use anti-aliasing filter.
3) Inverse FFT to Image Plane.
4) Apply Fresnel Pattern to plane.
5) FFT Back.
6) Repeat process for all planes.
7) Transfer plane back to w=0 plane. Output image.

The degree of correction for the w-term in Equation 3.2 is contingent on the spacing of the w-planes. A w-projection can also be applied to decrease the number of w-planes.

To validate that this mathematically works well, electric fields were simulated using Equation 3.3 for a series of point sources. Each source’s phase was randomised compared to the others to ensure they are incoherent. They were then sampled at a set of discrete locations in the measurement volume, corresponding to a 3D Gaussian distribution of points. This distribution is shown in Figure 3.1. Although no interferometer would look like this in practice, it allows

us to show that it is still possible to make correct images by applying w-Stacking to the electric fields, even in the artificial worst case.

The top image in Figure 3.2 is the sky brightness distribution without any w-correction applied. The image is completely incoherent and wholly unrepresentative of the true dirty map.

After w-correction using w-stacking, the sky brightness distribution shown in the bottom plot in Figure 3.2 is recovered. The stacks in this case are calculated every w=0.1. An anti-aliasing kernel is also applied to the image, using a prolate spheroidal wave function (Jackson et al. 1991). A difference image is formed versus an image calculated using the direct Fourier transform implementation of Equation 3.1, and this is shown in Figure 3.3. The dirty image is recovered to an error of 1 part in 10^4 averaged across the image, with pixels corresponding to points having slightly higher errors of between 1 part in 10^2 and 1 part in 10^3.

This demonstrates that wide field correction can be performed in the same way as visibilities using the electric fields, however this is not an efficient method in practice. w-Projection may be a good method with large grid sizes, but this is likely to not be practical until consumer compute hardware increases in power to allow an EPIC correlator implemented on a higher frequency interferometer with commensurate larger grid sizes, and the w-Projection overhead.

3.4 Direct Fourier Transform Imaging

Using a direct fourier transform is an attractive alternative mathematically, because of its perfect w-correction. However in practice this approach, especially with visibilities, is often computationally unfeasible. The approach was tested on datasets from the LWA to explore its performance, and a broader analysis was undertaken to understand the practicality of the approach, and its applicability to other arrays.

3.4.1 The Long Wavelength Array

The direct Fourier transform method was tested using pre-captured electric field data from the LWA. The original correlation pipeline described in Kent et al. (2019b) and Chapter 2, was modified to facilitate this. The original and modified pipeline were both implemented using the Bifrost framework (Cranmer et al. 2017)[a].

The LWA is an interferometer currently located at two sites in New Mexico, USA. The site used for our analysis here is the Sevilleta site, LWA-SV, which is the same one used for the demonstration of the EPIC correlator in Kent et al. (2019b). The LWA operates between frequencies of 10 and 88 MHz. Each site consists of 256 dual polarization dipole antennas with a wide beam. The array is organised into a dense central core of pseudo-randomly located

[a]The source code for the DFT pipeline is available alongside this thesis.

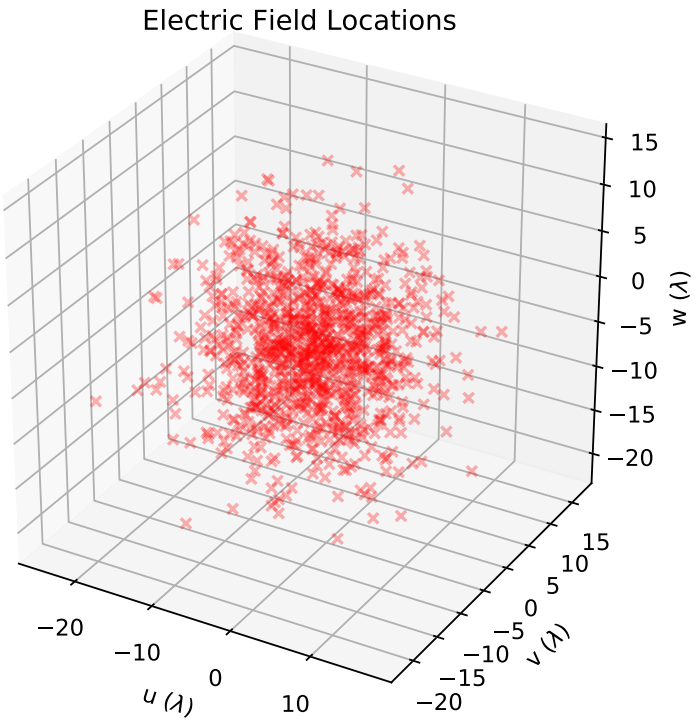

Figure 3.1: Simulated antenna locations for electric field measurement. Making a worst case, practically impossible distribution of points for co-planarity allows us to demonstrate that it can still be mathematically corrected.

antennas, with an outrigger antenna providing greater angular resolution. A more detailed overview of the instrument can be found in Chapter 2.

The LWA-SV antenna locations can be seen in Figure 3.4, where the color of each antenna marker corresponds to its w coordinate. The dense core is relatively flat with some minor non-coplanarity. The outrigger antenna, which greatly contributes to the overall angular resolution of the array, is several hundred metres away from the central core, with the antenna being roughly 10m higher than the rest of the array.

Originally, the LWA-SV EPIC correlator gridded the electric fields directly and then performed an inverse FFT to the electric-field sky space, followed by a squaring and accumulation operation to form the final image. This step was replaced by the DFT method described in Section 3.2.

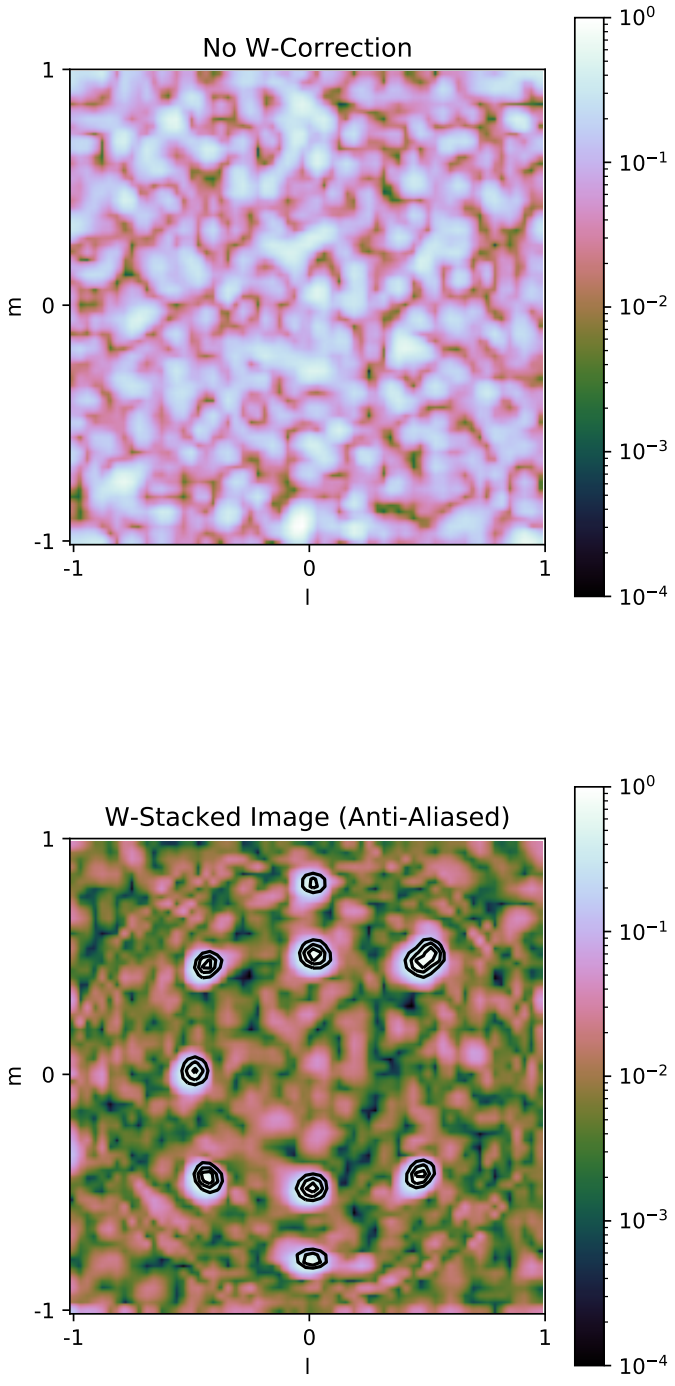

Figure 3.2: (top) The sky imaged without any w-correction. There is no realistic sky brightness distribution recovered. Pixel values are in arbitrary units. (bottom) A simulated sky imaged with w-correction using Direct W-Stacking. Our point sources are clearly imaged. Pixel values are in arbitrary units. The sky has been normalised to 1. Contours start at 0 and correspond to levels of 0.2. The colour scheme used is "cubehelix" (Green 2011).

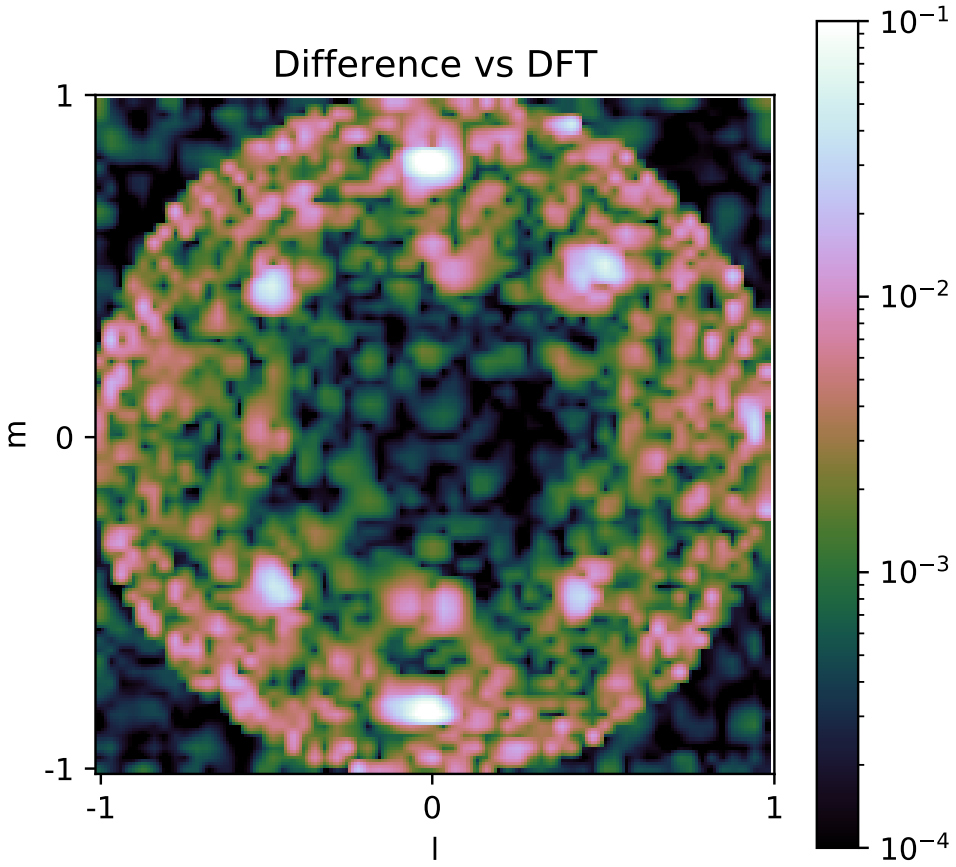

Figure 3.3: A difference image between the image in Figure 3.2 and an image created using the direct Fourier transform, with perfect w-correction. A PSWF anti-aliasing function is applied to the electric fields to account for sub-grid point sampling. Thus the equivalency of w-correction between E-Fields and visibilities is shown. However w-stacking with electric fields is not computationally feasible in real-time.

3.4.2 Validation with test Data

The DFT matrix was pre-calculated to not have any time dependence, and imaged sky cosine coordinates corresponding to a 64x64 grid in l,m space. This gives values both in and out of the celestial sphere, to simplify post-image rendering but constituting an "all-sky" image. The antenna cable delays and gains were factored into the DFT matrix using the LWA Software Library (Dowell et al. 2012).

The data used was captured at 74 MHz as Cygnus A and the Galactic plane transited

overhead. To demonstrate the wide-field errors that occur without w-correction, the DFT was calculated twice: once as stated above, and the second time with the w-term in the antenna locations set to zero. The latter method simulates imaging with no w-correction, similar to the original demonstration of the EPIC correlator. In an FFT based method, anti-aliasing of the antenna locations relative to the Fourier grid points also has to be accounted for, or else the error will increase further. With the DFT method, there is no need to compensate for aliasing.

A single channel image of Cygnus A and the Galactic plane is shown in Figure 3.5. An image difference between the DFT image and the image without w-correction is shown in Figure 3.6. The auto-correlations have not been removed, which add a DC offset to the image. The image is centred on the zenith directly overhead. In arrays with a constant slope (such as on a hillside) the effective zenith may not be directly overhead. The LWA is on such a slope, but still with significant non-coplanarity from a calculated best-fit slope. A phase correction can be factored in that corresponds to the geometry of this slope. In the case of the imaging here using the LWA this was not applied, and imaging is consequently centred on the zenith. Regardless, the majority of the non-coplanarity and wide-field error results from the outrigger.

As can be seen, even for an array such as LWA-SV, there are significant wide-field errors on the order of 10% for the dirty map produced. This is surprising, as it was originally thought that for a low frequency observation with an almost coplanar telescope such as the LWA site at Sevilleta, wide-field errors would be insignificant.

3.4.3 DFT Performance

The performance of using the DFT for this low frequency test case was found to be comparable to the original EPIC architecture of using a convolutional gridding followed by an FFT. This is likely due to the matrix multiplication used for the DFT, with dense matrix multiplications being particularly efficient when implemented on GPUs. Even with optimised algorithms such as the convolution scheme described in Romein (2012), and the CUDA FFT library, these algorithms do not map as well to the GPU model as do dense matrix multiplications; as a result, they suffer from low performance in comparison. The DFT multiplications were batched, with a single multiplication being performed for each channel and polarisation.

The run-times averaged over 50 correlation operations, for each time “gulp" of data is shown in Table 3.1. A time gulp in this instance corresponds to the coarse chunks of electric field data which are decimated in time that transit through the Bifrost framework. Each chunk consists of multiple timestamps outputted from the F-Engine. It was observed that a square data matrix, X, resulted in the most efficient multiplication times. This is likely due to the optimum benefits this provides in terms of locality, caching and sub-division of the matrix multiplication algorithm using a suitable “blocking" matrix multiplication algorithm (Lam

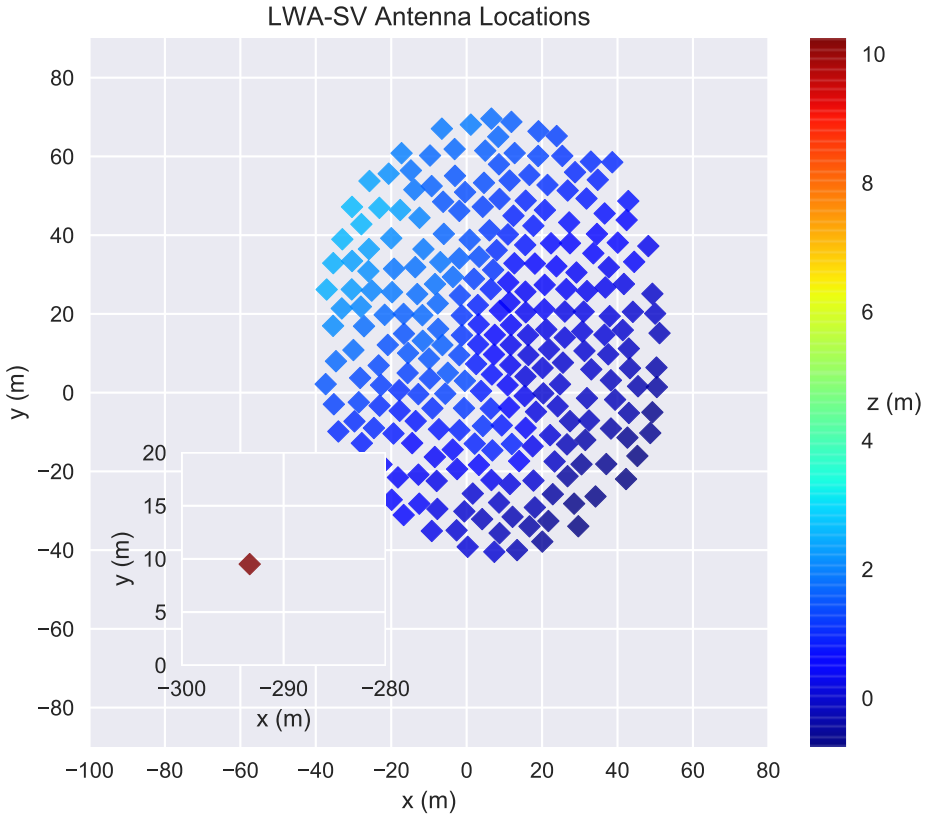

Figure 3.4: Antenna Locations of the LWA-SV. The w coordinate corresponds to the color of each antenna marker. As can be seen LWA-SV has some non-coplanarity within its dense core. The outrigger antenna, inset, is several hundred meters away and adds angular resolution to the interferometer, but is also at a greater w coordinate causing non-coplanarity.

et al. 1991). The DFT performs comparably to the original grid and FFT approach with EPIC. For perfect wide-field correction with the DFT it was possible to process 16 Channels of LWA data, corresponding to 400 KHz of bandwidth.

Figure 3.7 shows a simulation of the costs associated with doing direct imaging using the E-Fields with w-stacking versus the DFT or no w-correction at all. For a small number of antennas, w-stacking is very inefficient and the cost is high; this effect is made worse by the low operational intensity as shown in Figure 3.8, defined as the number of floating point operations performed on each byte loaded from memory, for the algorithms involved. All methods have approximately squared scaling as a function of the 1-D size of the 2-D grid. Thus this is unlikely to be practical for implementation on a radio interferometer with less than 50 elements.

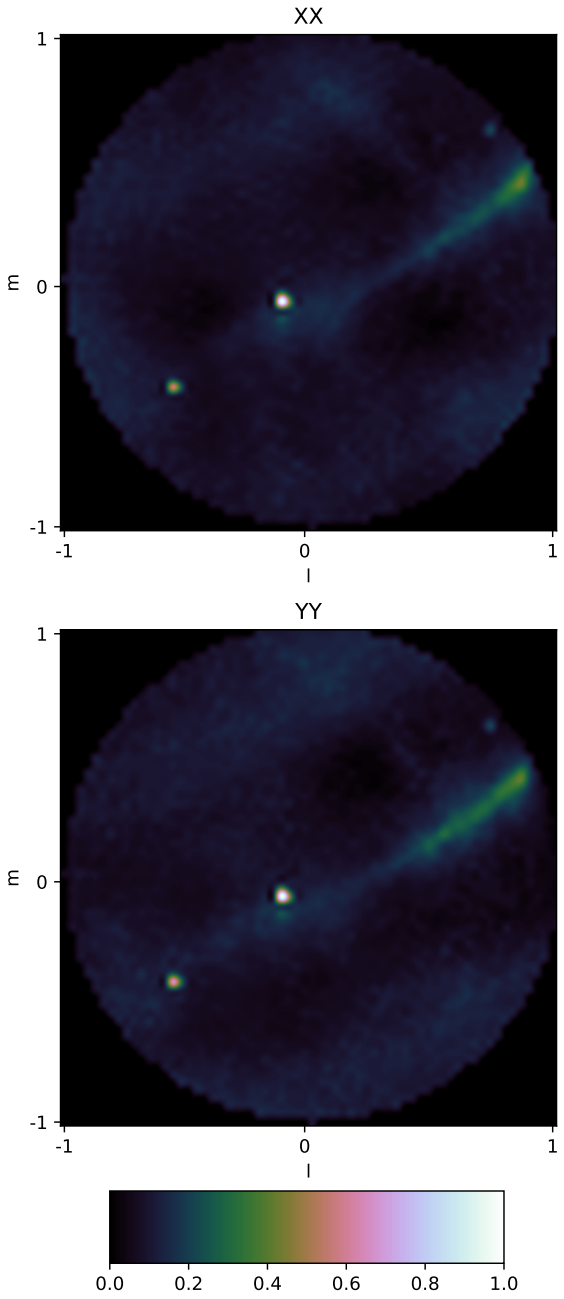

Figure 3.5: Image, normalised to the unity at the peak, formed with the DFT Matrix using the electric fields for low frequencies. This offers perfect w-correction within numerical precision. Auto-correlations have not been removed in this image, which gives an identical positive offset to all pixels. Cygnus A is the bright pixels in the central region of the image, with Cassiopeia A to towards the bottom left. The Galactic plane is also visible. Units are arbitrary and normalised to a max value of 1.

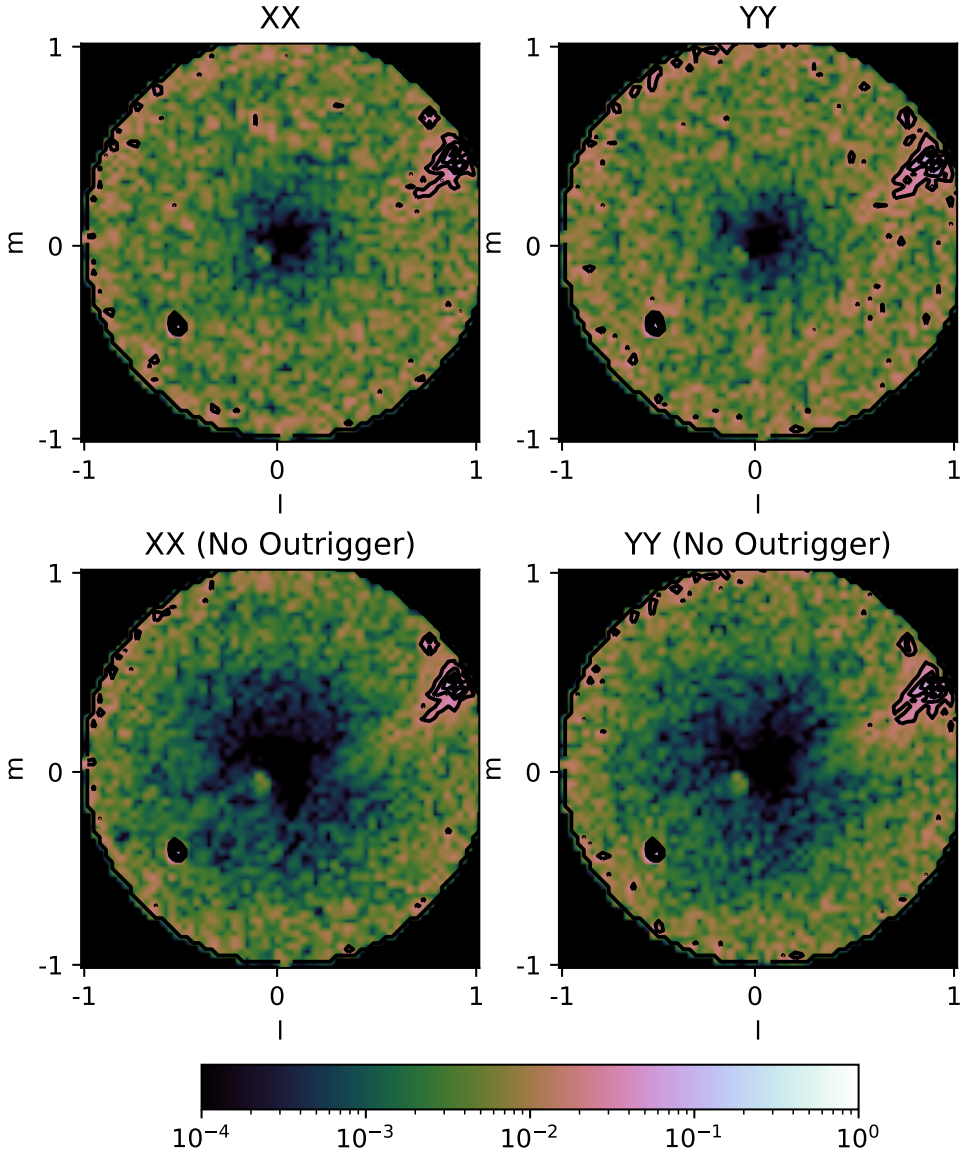

Figure 3.6: Difference between DFT Image and DFT Image with no w-Correction. This is equivalent to the difference between the DFT image and the image formed with the EPIC architecture where no wide field correction is made. Errors approach 10% low in the beam, with real astrophysical structure being strongly affected. Contours start at 0 and correspond to levels of 0.2. When the outriger is removed the wide field error is significantly less but still substantial, especially around structure such as the Galactic plane.

64^2 Image Size 8 Channels	Processing Time DFT	 EPIC
No. Timestamps		
512 (20ms)	11.8 ms	10.7 ms
1024 (40ms)	20 ms	13.3ms
16 Channels	DFT	EPIC
512 (20ms)	18.9 ms	11.621 ms
1024 (40ms)	40 ms	17.4 ms

Table 3.1: DFT Imaging run-times for a 64^2 image using a Pascal P100 GPU, implemented with the Bifrost framework, averaged over 50 runs. Bracketed numbers are the amount of time that a "gulp" of timestamps corresponds to. For a direct imaging pipeline to function normally the processing time must not exceed this figure; otherwise it is incapable of running in real time.

The DFT has strong performance in practice, when Figure 3.7 would lead us to believe it slower. To understand this, a roofline analysis has been performed (Williams et al. 2009). This is shown in Figure 3.8. The DFT benefits from significant operational intensity, which often translates into strong memory locality, with caching benefits, strongly benefiting run-time performance. Thus the algorithmic nature of the DFT is a good fit for modern GPU hardware, leading to higher real world performance than its raw floating point operation cost would lead us to believe. Compared to the original EPIC implementation, the DFT approach replaces both the gridding and FFT.

As the number of antennas increases however, a convolutional gridding and FFT approach may be more efficient provided the layout of the array is dense. A denser array allows smaller FFTs using the EPIC formalism (see Morales 2011; Thyagarajan et al. 2017; Kent et al. 2019b). However as shown, it is very challenging to perform wide-field correction in real-time using this approach.

3.4.4 DFT Applicability

The ability to use the DFT matrix as an architecture for accurate direct imaging is highly array and observation dependent. From the discussed costs and roofline performance figures, the DFT is very costly from a floating point operations perspective, but lends itself to very efficient implementations on current high performance compute hardware. Total floating point operations is a poor predictor of performance. The parameter that ultimately controls performance and applicability is the dimensions of the D matrix. Too many antennas, sky pixels, or a combination, can make this approach impractical for implementation on a real-time direct imaging interferometer. This is due to the computation time for a batch of electric field timestamps exceeding the amount of physical time each batch corresponds to.

The DFT architecture would be best applicable for a non-redundant and sparse array having many elements, where high time resolution imaging is a valuable observation mode. Redundant arrays can use FFT beamforming approaches (Masui et al. 2019) for reducing data rates, and dense arrays can use the EPIC correlator (Thyagarajan et al. 2017; Kent et al. 2019b) for direct radio imaging. A combination of these schemes can also be utilised, depending on the interferometer configuration, geometry, and science goals. It is important to remember that as an interferometer becomes more sparse, with implied increase in baseline length, ionospheric effects become more prominent. Correct characterisation of the ionosphere across the array is therefore extremely important for high fidelity imaging. These can be efficiently incorporated into a time-dependent D matrix but will increase the complexity of the interferometer.

An overview of current low frequency interferometers which may be applicable for this direct imaging technique are shown in Table 3.2. Apart from the LWA, HERA and the MWA are the most suited for the application of a DFT direct imaging technique as described here, due to the similar properties in terms of array topology and frequency coverage relative to the LWA. With current consumer GPU hardware, as used in the LWA-SV correlator, it would likely not be practical to deploy the described technique on CHIME or HIRAX. With the continuing rapid advances in consumer electronics, however, it may soon be a worthwhile technique.

In the example shown here, the resolution of the LWA-SV makes it possible to sample the entire sky using the DFT imaging approach at a resolution representative of the overall angular resolution of the instrument. If a greater resolution is required, such as for higher frequency measurements as with HIRAX or CHIME, or greater maximum baseline extent, the field of view may have to be restricted to allow usage of the DFT technique, or only tracking particular celestial objects. It may be that an EPIC or FX based correlator may be more efficient, dependent upon science goals. DFT matrix based imaging has the major advantage of being extremely flexible and can fit various interferometers for different observation modes.

With an array of high density, it may be more pragmatic to use the EPIC correlator architecture described in Thyagarajan et al. (2017), where gridding the electric fields onto a dense grid and then calculating the Fourier transform by the FFT algorithm is more effective. The EPIC architecture is not able to incorporate the outrigger antennas of the LWA due to them lying off the dense grid. Incorporating them would involve increasing the FFT grid size, resulting in a commensurate increase in computation time. The DFT can easily add outrigger antennas to improve instrumental angular resolution through increasing the diameter of the synthesised aperture by adding another column of the D matrix, with linear scaling as a result.

With this in mind, a direct imaging telescope using the DFT architecture described here provides accurate wide-field imaging, with no constraints on extent of the interferometer. With the addition of a calibration architecture as described in Beardsley et al. (2017), it is a highly

Telescope	Array Bandwidth	Channel Bandwidth	Antenna Elements	Stated FoV	Angular Resolution
LWA-SV (Taylor et al. 2012)	10 MHz - 88 MHz	100 kHz	256 Dipole Antennas	2.2° (74 MHz)	192′(74 MHz, Single Station)
HERA (DeBoer et al. 2017)	50 MHz - 280 MHz	97.8 kHz	380 14m Parabolic Dishes	9°	5.14′ - 25.2′
MWA (Tingay et al. 2013; Wayth et al. 2018)	80 - 300 MHz (30.72 MHz processable)	40 kHz	256 Tiles of 4x4 Dipole elements	24.° (150 MHz) - 19.4° (200 MHz)	2′(Precursor) - 3′(Full Array)
CHIME (Bandura et al. 2014)	400 MHz - 800 MHz	390 kHZ	1280 feeds across 10 cylinders	100°(EW) 2.5° - 1.3°(N-S)	30′ - 60′
HIRAX (Newburgh et al. 2016)	400 MHz - 800 MHz	390 kHz	1024 6m Parabolic Antennas	7.48° - 3.87°	5′ - 10′

Table 3.2: An overview of current low frequency radio interfereometers that may be suitable for application of DFT imaging using the electric fields. The LWA-SV (tested), HERA and the MWA are the most applicable targets. HIRAX and CHIME work at higher frequencies necessitating sampling a high number of sky points to form an image representative of the angular resolution of the instrument. They also have more antenna elements, which along with the higher frequency, might make the DFT imaging approach described in this paper practically unfeasible with current compute hardware. In the future, with the continuous advance of consumer electronics, this may change.

accurate, fast, and flexible method for direct electric field imaging.

3.5 Discussion

Use of a Direct Fourier Transform imaging matrix has been shown to be a tractable solution to the problem of wide-field imaging on direct electric field imaging interferometers, especially at low frequencies. This is only possible by imaging the electric fields directly from each antenna. A similar approach using visibility data products would be practically unfeasible due to the high computational load.

As has been shown, standard techniques for wide-field imaging with visibilities are a poor fit to imaging directly with the electric field data from each antenna, due to the requirement for them to be done every time step in real-time. The compute cost in floating point operations is made worse by the nature of the algorithms involved. FFTs perform poorly on GPUs, especially at small sizes, and the difficulty in predicting the memory access pattern of convolutional

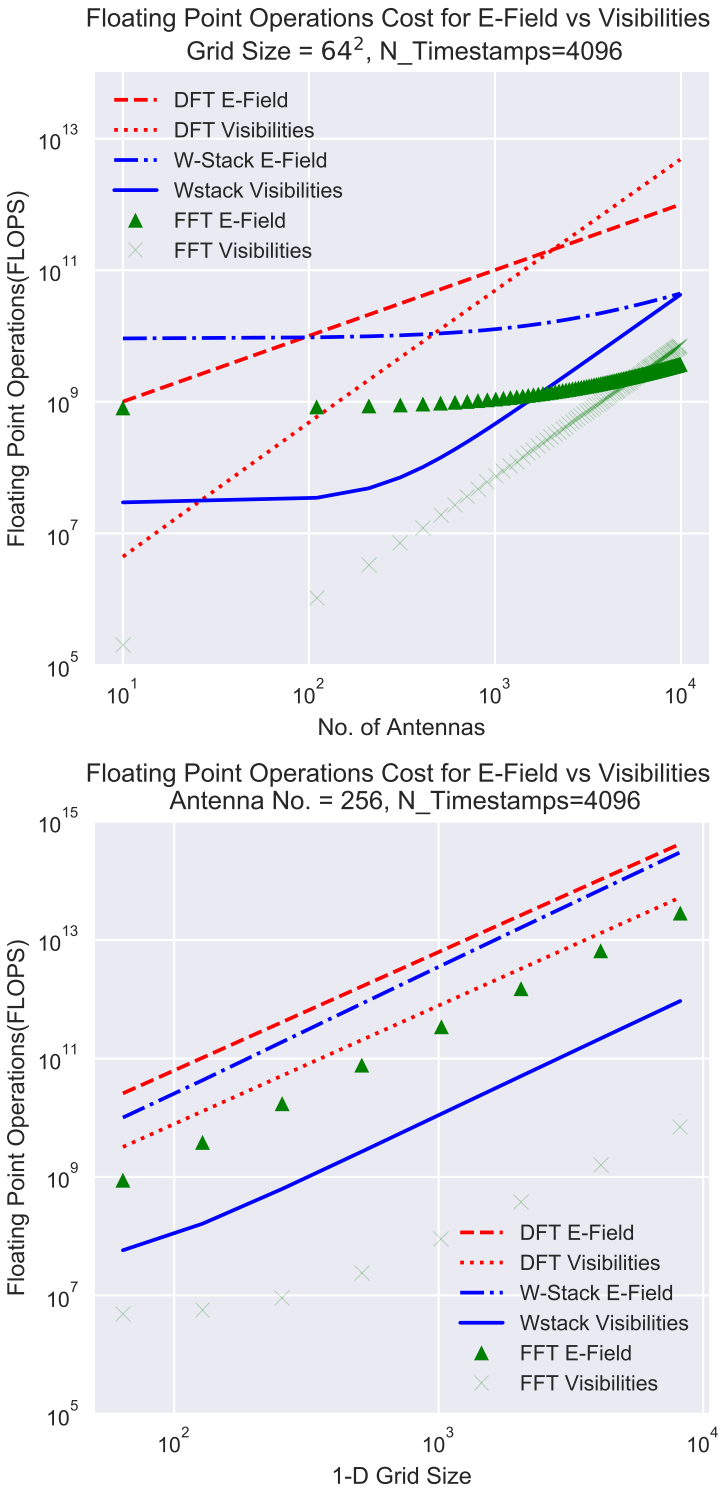

Figure 3.7: Naive Floating point costs calculated for different wide field imaging techniques as well as the cost of the "standard" EPIC architecture using convolutional gridding and an FFT. Whilst the DFT cost is often the highest, algorithmically it lends itself to efficient computational implementation which can negate this in certain parts of the parameter space, and have comparable efficiency to convolutional gridding and FFT based direct imaging.

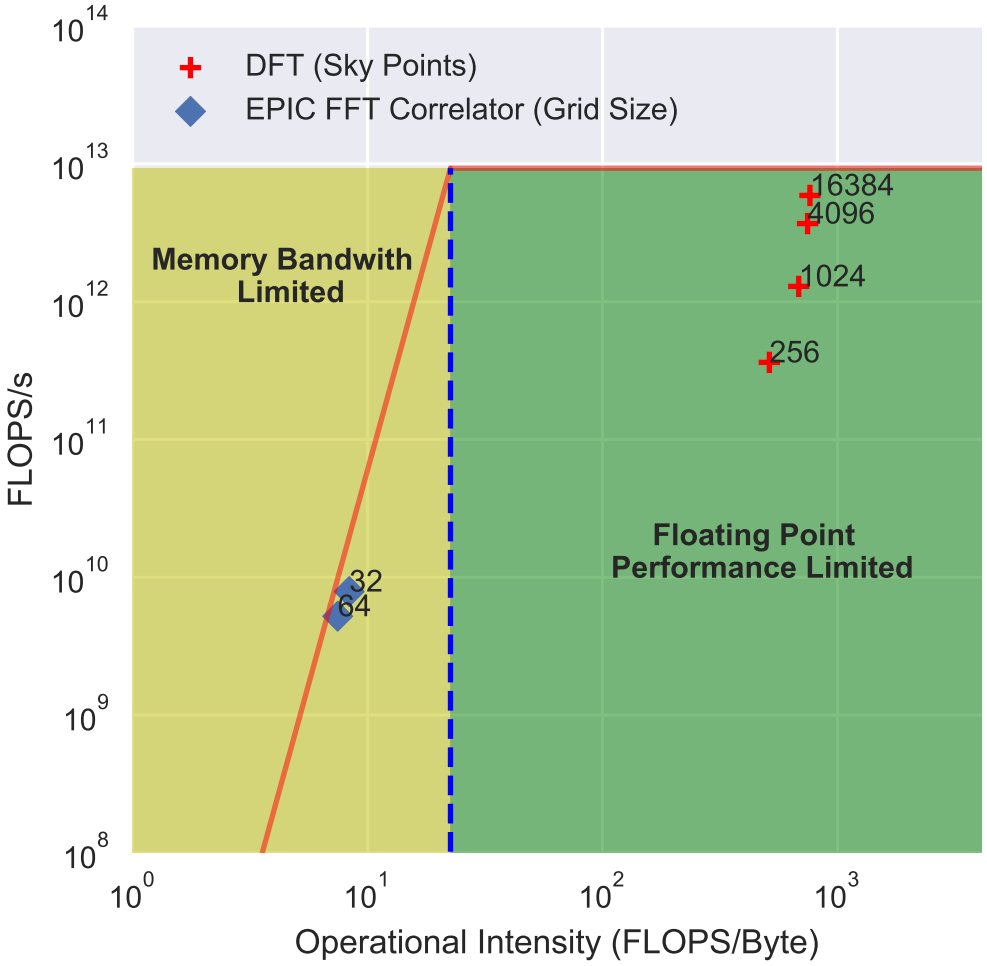

Figure 3.8: Roofline analysis of the DFT method versus the EPIC FFT pipeline. The roofline plotted is for an NVIDIA Pascal P100 GPU. The DFT benefits from much greater operational intensity compared to FFT based methods, which is the reason it runs at comparable speeds despite having significantly more operations. Greater operational intensity allows locality to be taken full advantage of.

gridding also decreases efficiency.

The DFT based approach places no restrictions on the placement of antennas or on regularly spaced grid points, unlike with an FFT. The scaling of this method is $O(N_K N_T N_A)$, which means that adding in additional antennas, sky pixels, or more timestamps causes a linear increase in cost scaling.

Thus the DFT allows real-time high time resolution wide-field imaging, with significantly enhanced flexibility compared to gridding and FFT based methods. The D matrix described

in Section 3.2 allows important antenna dependent terms to be taken into account. Further extension of the formalism to include direction-dependent effects such as ionospheric distortion per antenna is also mathematically simple, but we do not have access to a suitable instrument to test this with currently.

Use of a DFT matrix allows the highly compute bound nature of matrix multiplication to be taken advantage of, and GPUs are exceptionally efficient for compute-bound algorithms of this type. The ability to define sky points to sample at will opens the possibilities of variable resolutions across the dirty map. Adding time-dependence to the D operator allows calibration terms to be additionally taken into account to create a full featured interferometric imaging framework.

The time-dependence of the D operator also makes this formalism a profoundly powerful technique for transient radio science, with its ability to track sources across the sky with very precise placement of each imaging pixel, To the best of our knowledge, there is not another radio imaging technique that is able to image with such high temporal resolution and spatial resolution. Spatial resolution is limited by the instrument itself, namely the longest baselines.

From the demonstration of this direct imaging technique, a corollary follows: the total floating point cost is an extremely blunt instrument and provides a profoundly flawed way of predicting algorithmic runtime. Conventional wisdom in radio astronomy has a high regard for this metric; in contrast, thinking in terms of a roofline model, using the computational cost in concert with the available memory bandwidth is a small step in the right direction.

The most important conclusion here is that the mathematics of the van-Cittert Zernike theorem in Equations 1.17 and 1.16 can be re-arranged in several different ways to get the same physical quantity, $I(l,m,n)$. The computational implementation of each mathematical sleight of hand for Equation 1.17 that we have used in the previous two chapters are vastly different, however. The wisdom derived here is that thinking of a mathematical relation in only one form will constrain your ability to make a computationally efficient implementation. Re-arranging Equation 1.17 into the DFT linear operator formalism in Equation 3.5, with the same underlying semantics, is very well suited to a GPU.

Thus if an algorithmic implementation on a particular piece of compute hardware seems inefficient then perhaps, with a good grasp of the underlying mathematics of the problem at hand, that problem can be reformulated into something that is very efficient, such as that discussed in this chapter, even when conventional wisdom and metrics would describe it as inefficient. There are an alarming number of cases of dreadfully unsuitable algorithms being implemented on whatever is the most fashionable piece of computing hardware at the time.

Chapter 4

Redundancy Analysis of the Hydrogen Epoch of Reionisation Array using the Bispectrum

4.1 Introduction

The Epoch of Reionization is the last great phase change of the universe, where the neutral intergalactic medium (IGM) is ionized by hard radiation background emitted by the first stars and active galactic nuclei (AGN). From a wide variety of simulations on different scales, the 21cm line has emerged as one of the strongest probes of the evolution of reionization at different red-shifts (Mesinger 2017; Mesinger et al. 2011). The red-shifted signal provides a strong probe at all stages of the reionisation process.

Instrumentally, this wavelength can be detected by low frequency radio interferometers. The low frequency foregrounds are significantly stronger than the HI emission, however. The foregrounds are dominated by strong continuum's due to synchotron emission, of order 10,000 times stronger than the HI emission. Examples of capable interferometers include HERA, HIRAX, CHIME and the SKA (DeBoer et al. 2017; Bandura et al. 2014; Dewdney et al. 2009a).

Different approaches have been posited for uncovering the HI signal which can be divided into Foreground Removal and Foreground Avoidance techniques. Foreground removal is based around the idea of removing all point sources and continuum emission within the field of view of interest for observing the HI signal. In practice this is exceptionally difficult, with extremely

accurate and precise sky models needed for this method to prevail.

Another approach is to use Foreground Avoidance techniques, where the smooth foreground emission can be separated from the rapidly oscillating HI emission. One of the leading methods for this is the Delay Spectrum technique (Parsons et al. 2012). The delay spectrum technique requires highly accurate calibration schemes to achieve a thermal-noise limited delay spectrum, where the noise can be reduced by averaging. Highly redundant arrays can make use of redundant averaging based on the fact many different baselines of the same length have the same response to the sky, so long as they are redundant with each other, with identical beams (Li et al. 2018; Zheng et al. 2014).

Recently, there has been discussion of using the closure phases of identical triads as a method of detecting the HI signal using the bispectrum (Thyagarajan et al. 2018), alongside the delay spectrum technique. The closure phases have a long history in radio astronomy, following their discovery by Jennison (Jennison 1958). This is where three antennas whose baselines form a closed triangle form a single observable phase, which is free of direction-independent gains from each antenna. The closure phase is therefore a direct observable of the sky with thermal noise imposed on top as discussed in Carilli et al. (2018), but without need for calibration. The closure phases, and the related closure amplitudes, also form the lynchpin of several imaging and calibration strategies, as mentioned in Chapter 1.

The closure phase method is intended for use in an array with a redundant layout, such as HERA, HIRAX, or CHIME, which allows plenty of triangles of identical shape and orientation to be formed between the different antennas in the array. Identical triads of the same orientation should have the same closure phase, if they are completely redundant. If the phases within each set of triads depart from each other, this implies a non-redundancy due to either direction-dependent effects, such as beam errors, or some other form of closure error.

As discussed in Carilli et al. (2018), closure phases also provide a unique method of probing non-redundancies in arrays with a redundant layout. Here we present a development of this technique along with a detailed analysis of non-redundancies in the HERA array. The usage of circular statistics is formalised to provide an overall redundancy 'map' of the HERA array for a particular triad group as a function of Local Sidereal Time and Frequency. The frequency range investigated is from 100MHz to 200 Mhz. In addition to this, modeling is performed of a redundant layout telescope such as HERA, to determine which direction dependent effects are most significant for disrupting redundancy between triads. The modeling is informed by direct electromagnetic simulations (Fagnoni et al. 2019).

An overview of the different theoretical treatments used in our analysis are summarised in Section 4.2. The dataset and HERA are discussed in Section 4.3. Our modeling efforts, so as to shed light on the sources of non-redundancy in a redundantly laid out telescope using the

closure phase is discussed in Section 4.4. The results of our redundancy analysis on the data from the HERA telescope, and comparisons to our modelling efforts are shown in Section 4.5. Finally a discussion of the current limitations that need to be overcome, both by the closure phase methods and parallel visiblity methods, along with implications for future development of HERA and reionisation experiments, is discussed in Section 4.6.

4.2 Theory

The closure phase, C_{abc} is calculated between three antennas a,b,c over the frequency range of the telescope, from their visibilities, V_a,V_b,V_c. The visibility relation is calculated as defined by Equation 1.16, with the closure phase derived as per Equation 1.37. This approach has the benefit of canceling all instrumental direction-independent gain terms:

$$C_{abc}(\nu) = \phi^s_{ab} + (\theta_a - \theta_b) + \phi^s_{bc} + (\theta_b - \theta_c) + \phi^s_{ca} + (\theta_c - \theta_a) + \phi^n_{abc} \tag{4.1}$$

In the presence of a strong point source, the closure phase responses of each baseline will cancel with each other, and the closure phase will approach zero. More complicated sky structures will result in a departure from zero, and an overall more complex spectrum of closure phases when viewed as a function of frequency. This spectrum is known as the 'bispectrum'. Fuller details are given in Carilli et al. (2018).

Using Closure Phases as a measure of Redundancy

Beginning with the assumption that the closure phase measured on the same triangle at the same LST on different days differs only by circular Gaussian random noise as per the Wrapped Normal Distribution (Mardia and Jupp 2008), so that:

$$\phi_{a,b,c}(\nu, t; D) = \hat{\phi}_{a,b,c}(\nu, t) + \epsilon \tag{4.2}$$

where $\phi_{a,b,c}(\nu, t; D)$ is the closure phase measured on triangle a, b, c at frequency ν and local sidereal time t and date D, $\hat{\phi}_{a,b,c}(\nu, t)$ is the underlying true closure phase and $\epsilon \sim WN(0, \sigma_\nabla)$ where σ_∇ is the angular noise on a triangle's phase. We can estimate the noise σ_∇ by calculating the standard deviation of $\phi_{a,b,c}(\nu, t; D)$ across different days assuming the closure phase across different days differs only by noise.

The standard deviation σ_∇ is calculated as Mardia and Jupp (2008) by first wrapping the phases around the unit circle: $cos(\phi) + isin(\phi) = e^{i\phi}$. The mean resultant vector inside the unit disk is calculated:

$$\frac{1}{N}\sum_{x=0}^{N} cos(\phi_x) + \frac{i}{N}\sum_{x=0}^{N} sin(\phi_x) = \bar{R}e^{i\bar{\phi}} \tag{4.3}$$

where ϕ_x represents each closure measurement at a particular $\phi_{a,b,c}(\nu,t;D)$, thus N measurements, such that $x = 1,2,..,N$. The set of measurements of size N is the set of all triads of the same size and orientation. As $0 < \bar{R} < 1$ we can use it as a measure of how concentrated are the closure measurements at a particular channel at a particular timestamp. If they are highly concentrated ($R \sim 1$) then the individual closure measurements in each triad class are very similar and redundancy is good. If $R << 1$ then redundancy is poor. This parameter is a direct measure of redundancy. The characteristic function of the wrapped Gaussian is:

$$\int_0^{2\pi} e^{in\phi} f_{WN}(\phi;\bar{\phi},\sigma)d\phi = e^{in\bar{\phi}-\frac{n^2\sigma^2}{2}} \tag{4.4}$$

Where f_{WN} is the characteristic function of the wrapped normal distribution. Thus we arrive at Equation 4.3 considering integer values of n:

$$e^{i\bar{\phi}-\frac{\sigma^2}{2}} = \bar{R}e^{i\bar{\phi}} \tag{4.5}$$

From $\bar{R} = e^{-\frac{\sigma^2}{2}}$ we can then calculate standard deviation on the circle as $\sigma_\nabla = \sqrt{-2\ln(\bar{R})}$. The mean closure phase is calculated as:

$$\bar{\phi}_\nabla(v,t;D) = \tan^{-1}\left(\frac{\bar{S}}{\bar{C}}\right) \tag{4.6}$$

where $\bar{S}$ and $\bar{C}$ are the sine and cosine terms of $\bar{R}e^{i\bar{\phi}}$. $\bar{\phi}_\nabla$ represents the mean closure phase of a closure triad at a particular frequency and LST. From this, we can find the average of all triads on a single LST, and then average those averages over successive LST's to reduce the noise. The set of averages, with each element of the set denoting a triad on a single LST, can be used for finding the standard deviation as above.

4.3 Hydrogen Epoch of Reionisation Array

The Hydrogen Epoch of Reionization Array (HERA) is a radio interferometer located in the Karoo Desert in South Africa (DeBoer et al. 2017). In its final configuration, it will consist of 350 individual antenna element in a redundantly arranged geometry. There will be a dense core of 320 antennas, with a further 30 antennas acting as outriggers to increase the angular resolution of the instrument. This redundancy is a cornerstone characteristic in that it allows the usage of redundant calibration, such as that discussed in Li et al. (2018); Thompson et al. (2017); Zheng et al. (2014). The antennas are paraboloid in shape, with locations carefully constrained to within 10cm of their nominal positions.

HERA operates between 100 and 200 MHz, which will be increased to 50 MHz to 250 MHz with the installation of a new correlator. With the correlator used to generate the dataset

in this analysis, there are 1024 channels, with a channel bandwidth of just under 100 kHz. With the bandpass filtering taken into account the effective band lies between 110MHz and 190MHz. The visibilities generated are averaged over 10.7s, generating a single session.

The dataset used in the present analysis is the HERA IDR2 dataset (Dillon 2018) which was generated between, and including, the days JD2458098 (10th December 2017) to JD2458118 (30th December 2017). The data consists of raw, uncalibrated visibilities from which closure phases are generated for different sets of identical triads. The correlator outputs sixty sessions at a time and stores them in a MIRIAD file format; these are converted to the CASA Measurement Set (ms) format, using pyuvdata (Hazelton et al. 2017), through the intermediate data format of UVFITS.

The closure triad classes used here comprise two equilateral triad classes and three linear triad classes. The equilateral classes consist of Equilateral 14m baselines (EQ14), and Equilateral 28m baselines (EQ28), both oriented North. The linear classes consist of a triad with two 14m baselines and a 28m baseline, oriented East-West(LinEW), 30 degrees to the North (LinM30) and -30 degrees to the North (LinM30). A figure showing the redundant layout of the array can be seen in Figure 4.1.

Closure phases are generated individually for each miriad file outputted by the correlator, which form a data cube of shape [Sessions, Triads, Channels]. With our five triad classes, there are thus generated five different closure files for each input miriad file. The HERA correlator only observes between times fixed in the solar time frame of reference, thus the sky drifts by four minutes per solar day when measured in sidereal days using Local Sidereal Time (LST).

LST alignment was done at the same time as a closure phase data "cube" is generated. The closures are extracted from their individual files from each correlator output, stitched together into a day long array, and then aligned the correct amount relative to each other to account for the drift in sidereal day versus the solar day. The LST ranges of interest are then extracted, and a data cube of closure phases generated of shape [Sessions, Days, Triad, Channels]. This data pipeline is closely tied to that for generating a cosmological delay spectrum using the bispectrum, and is covered in greater detail in Chapter 5.

4.4 Redundancy Simulation

The notion of redundant arrays is based around the idea that each baseline in a redundant group should have an identical response to the sky, to within some percentage. This idea translates into the closure phase methodology as follows: geometrically identical and identically oriented triads should measure almost exactly the same closure phase, with phases in a set closely clustered with a small standard deviation, σ for the sample. There will be a scatter between

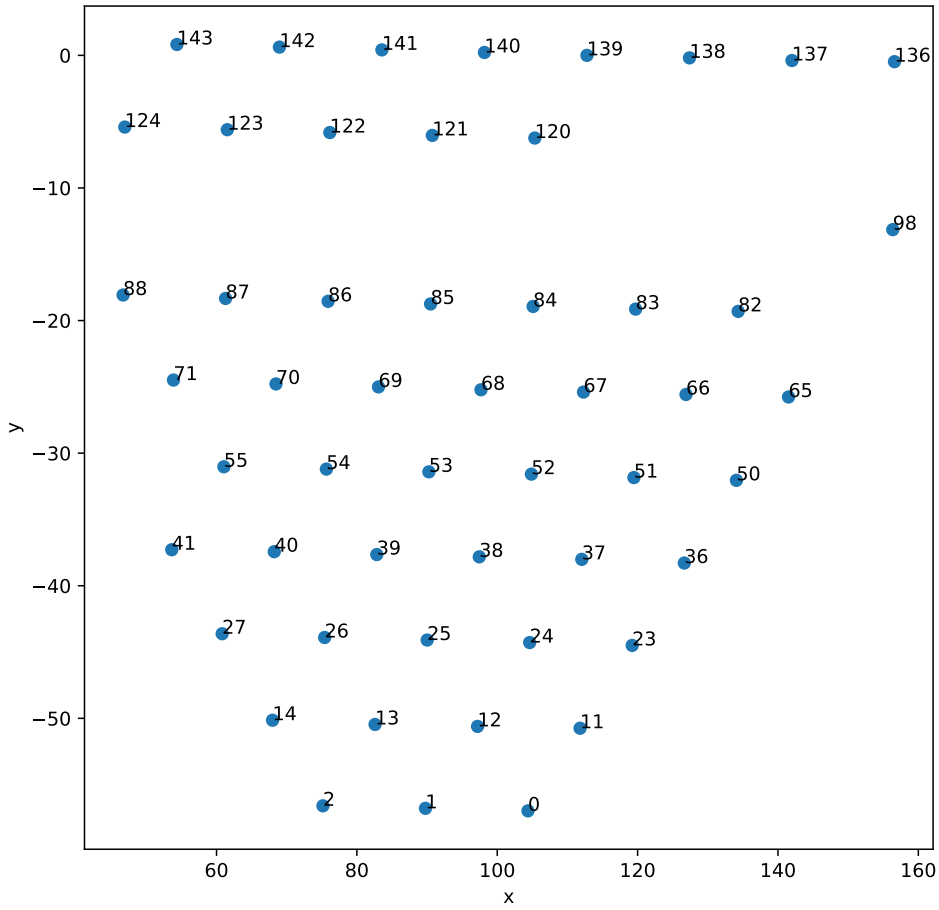

Figure 4.1: HERA Antennas operational during the IDR2 dataset. Coordinates are in the local orthogonal co-ordinate system used by the array correlator.

the closure phases inherently from the noise on each visibility measurement, with the scatter a product of the signal-to-noise ratio, as shown in Appendix A.

As closure phases are immune to direction independent antenna effects by definition, any deviation between redundant triads can only be the result of the telescopes non redundancies. These can include direction-dependent antenna effects such as beam differences, or closure errors due to irregularities in correlator electronics and errors in antenna location. Simulations were made (see below) to attempt to model a basic HERA-like array geometry to try and narrow down what can cause a divergence in closure phases indicative of a potential non-redundancy in the telescope.

By modelling deviations between closure phases using circular statistics, we can see how this redundancy varies across frequency and LST. Redundant arrays facilitate insightful closure diagnostics regarding array performance, which can help end users understand insidious non-redundancy effects which are liable to complicate data analysis pipelines.

To investigate the possible causes of non-redundancy, some simplified modelling was made of an interferometer of a deliberately redundant layout, such as HERA. The antenna locations and baselines are shown in Figure 4.5. The aim of these investigations was to model a simple sky of point sources, and vary the beam shapes and correlated baseline noise so as to quantify the effect on the closure redundancy. The beams were randomised and the error between them calculated as the Mean Absolute Difference (MAD) between the average of the antenna beams and each individual beam:

$$MAD = \frac{1}{l}\frac{1}{m}\sum_{0}^{l}\sum_{0}^{m} F(l,m)_a - F(l,m)_b \tag{4.7}$$

The closures were then generated for a sky modeled using the GLEAM catalogue at 143 MHz, without continuum emission. Six additional sources were added, corresponding to some of the strongest radio sources in the sky, which are excluded from the GLEAM catalogue; these are Fornax A (NGC 1316), Cassiopeia A, Cygnus A, Centaurus A, Messier 87 and Taurus A. At a particular timestamp t, this leads to a set of sources in the sky:

$$\bar{K}_t = \{\bar{k}_0, \bar{k}_1, \ldots, \bar{k}_{N_k}\}_t \tag{4.8}$$

A threshold was set for this sky model so that only sources above 8 Jy in strength were included, to make the simulation more computationally efficient. The set of sources becomes:

$$K_t = \bar{K}_t > 8Jy = \{k_0, k_1, \ldots, k_{N_k}\}_t \tag{4.9}$$

where K_t is the set of sources above 8Jy in strength at timestamp t. Each source has an associated set of l, m coordinates. This resulted in an average of 380 sources above the horizon at each LST time, t, that was sampled. After this, the visibilities are generated for each baseline in the array layout:

$$V_{ab,t} = \sum_{k=0}^{N_k} I_k(l_k, m_k, n_k) F_i(l,m) F_j(l,m) \exp\left[-2\pi i(l_k u_{ab} + m_k v_{ab} + n_k w_{ab})\right] \tag{4.10}$$

where a and b represent indices in a set of N antennas, thus Equation 4.10 specifies a visibility matrix. F is the direction dependent antenna beam, also known as the 'primary beam'. From this the closure phases can be calculated between three baselines from three antennas which form a closed triangle as is specified in Equation 1.37.

The redundancy statistics were then formed in a set of closure phases which correspond to identically shaped and oriented closure triads, using the method discussed in Section 4.2. This then allows redundancy to be characterised against a realistic sky and beam model, with the caveat that diffuse emission is not included. The redundancy as measured by the deviation in the closure phase between triads should not depend strongly on the presumably smooth diffuse component, but on the direction dependent non-redundancies from the beam.

4.5 Results

Measurements of the redundancy and the resulting delay spectra are presented for the HERA IDR2 dataset. The observations of redundancy are complex, with patches of strong redundancy as quantified by the spread between identical geometric triads, and large areas of poor redundancy.

These direct measurements of the HERA closure redundancy are presented alongside simulations of array parameters and how these translate to the redundancy of the closure phases. The closure phases are quite sensitive to small deviations in the power beams of each individual antenna, as well as to correlated baseline noise, which is of concern in a closely packed array such as HERA.

4.5.1 Redundancy of the HERA Array

The Closure redundancy for a triad class that consists of 48 14m Equilateral triads is shown in Figure 4.2. The transit of Fornax-A corresponds to the cold patch between LST of 3 and 4. In Carilli et al. (2018) it was noted that in the presence of a strong point source, the closure phases will be zero. In the case of Fornax A it is almost a point source, as it is slightly resolved relative to HERA's dirty beam, and dominates the primary beam response of the telescope during transit.

Outside of this field, the sky structure is less trivial which excites non redundancies as measured in the closure phases, and by abduction, the visibilities.

Upon looking across successive days it becomes clear that for each triad class, the plot as seen in Figure 4.2 is repeated with almost no deviation apart from clearly errant days. Each plot shifts by 4 minutes to the left due to the drift of sidereal time relative to solar time. This extraordinary repeatability across days indicates that the non-redundancy we see can only be excited by the sky. Different classes of triads exhibit different characteristics of non-redundancies which results from different fringe sampling of the sky from each triads baselines. Figures 4.3 and 4.4 show the redundancy on JD2458098 for 31 Equilateral 28m triads, and 26 Linear East West triads respectively. Equilateral 28m triads compose an equilateral triangle

of 28m baseline lengths, and Linear East West triads comprise two 14m baselines and a 28m baseline. Horizontal lines, consistent through most or all LST records, are radio frequency interference.

4.5.2 Redundancy Modelling of the HERA Array

The effects of the differences in beam shapes is shown in Figure 4.6. For a simulated sky of point sources, with Gaussian's of randomised widths, the resultant closure phases vary dramatically. Even errors as small as 1%, as measured by the Mean Absolute Difference (MAD) are enough to cause significant divergence. As this non-redundancy increases, it will cause the direction dependent complex gain response to the HI signal to change, causing decorrelation of the signal over redundant groups of baselines and triads.

Correlated baseline noise, which could result from an effect such as a standing wave or other correlated signal on the telescope can cause divergence too. Nuance is required in interpreting this, as the HI signal will also show up as a correlated noise-like signal on each baseline, however the HI component will be extremely small. Simulated closure redundancies for different ratios of correlated additive noises are shown in Figure 4.7. As can be shown, a correlated term breaks the closure relationship and causes a non-redundancy in the resulting closure spectra, and by abduction, the visibility spectra.

It has been noted by (Fagnoni et al. 2019) that in simulations of a HERA-like array with idealised parabolic reflectors and feeds there is a strong coupling effect between the antennas when the array is excited by a plane wave incident upon the array. This both shifts all of the beams to be not unique, breaking antenna beam redundancy, and propagates through the array as a correlated signal.

To understand whether this could contribute to non redundancy between identical triads, a model array with a redundant geometry was simulated with a sky model of GLEAM sources, as discussed above. The results of this simulation for different errors between the beams of each antenna is shown in Figure 4.8. When using an Airy disc approximation for each antenna beam, it can be seen that the sidelobe and primary beam gain errors have a negligible impact on the redundancy in this model. The most prominent cause of error is a pointing error of 1 degree. A 1 degree pointing error is in agreement with those predicted by electromagnetic simulations and on site estimates of HERA feed locations relative to the parabola (Fagnoni, Private Communication).

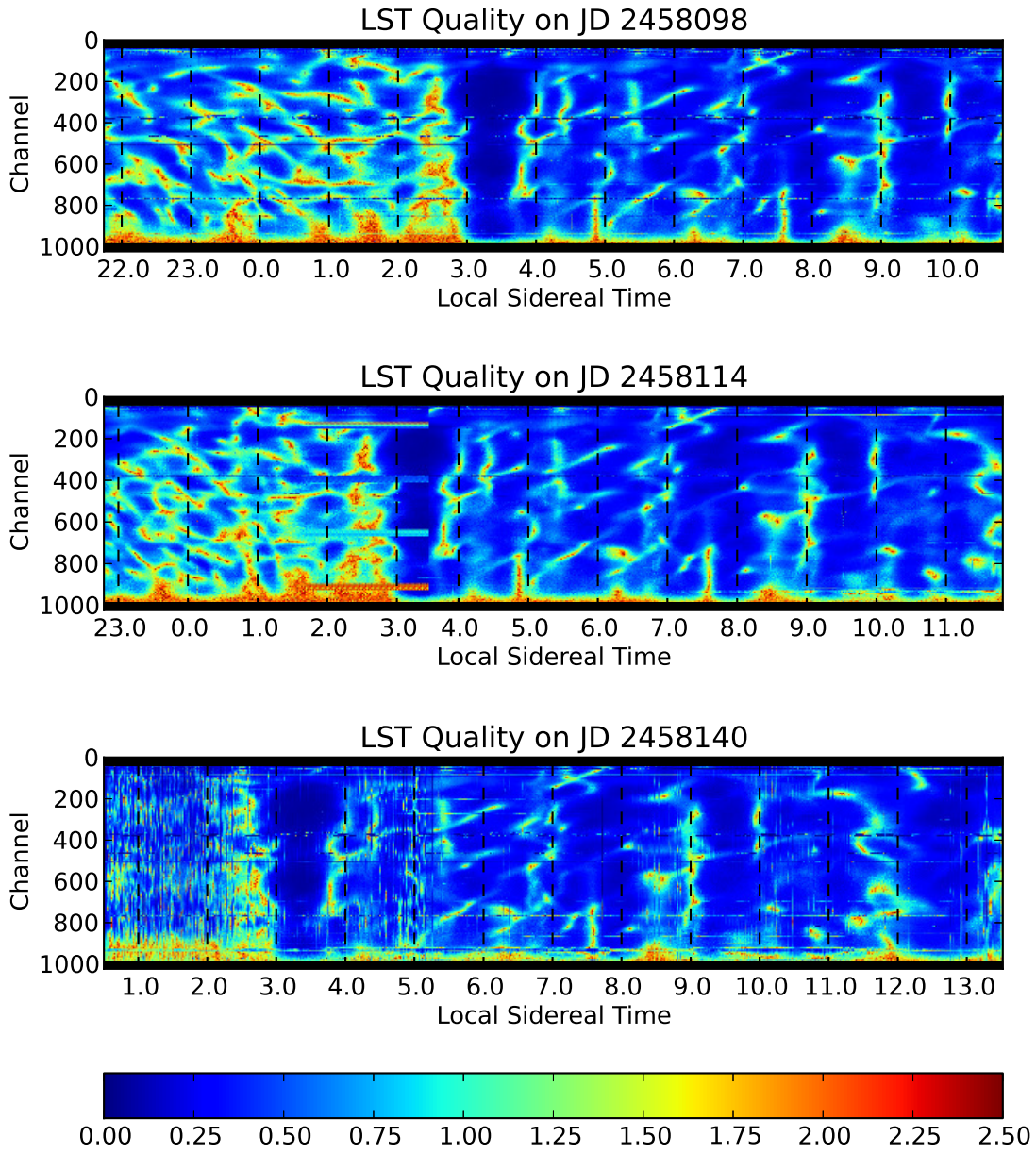

Figure 4.2: A plot for Equilateral 14m X Polarisation triads showing the standard deviations across the closure phases for each triads, plotted for Local Sidereal Time vs Channel. The colour scale is in units of standard deviation of the phases, at a certain sidereal time and frequency channel. The deviations have a unique pattern when viewed in this way. The deviations are identical at the same LST from day to day. Thus the non-redundancy is a property of the instrument excited by the sky.

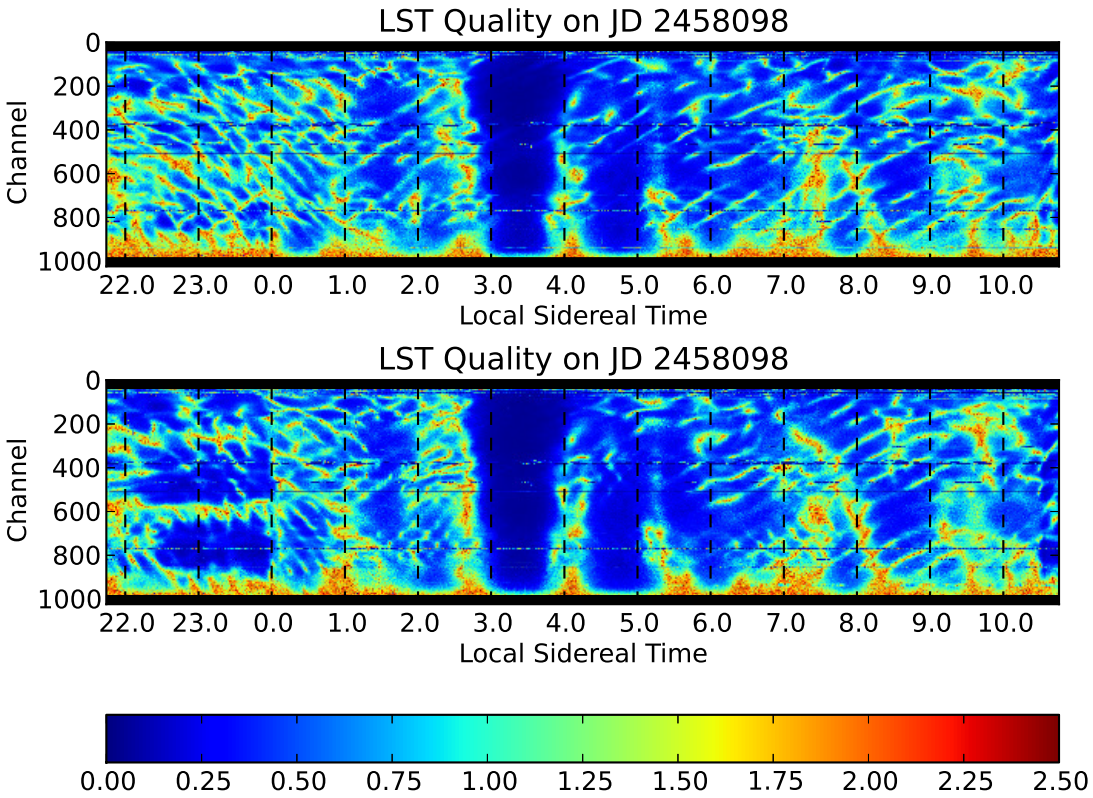

Figure 4.3: Redundancy for Equilateral 28m X (top) and Y (bottom) polarisations for JD 2458114. The colour scale is in units of standard deviation of the phases, at a certain sidereal time and frequency channel.

4.6 Discussion

Use of the closure phases for diagnosing the redundancy of the HERA array is demonstrated, extending work in Carilli et al. (2018) by making a more rigorous use of statistics and modelling on data from the entirety of the HERA IDR2 release. The major benefit of the closure phase is that its calibration-independent nature means that divergence between triads in a redundant grouping can only be a result of either direction-dependent effects, or some other form of closure error.

We see systematic divergences between triads as a function of LST and Frequency at the same time every day for different groups of redundant triads. This extraordinary repeatability implies that these divergences are a sky excited response on each baseline in a closure triad, meaning in turn that closure errors from correlated noises or other terms are unlikely to be a primary driver as these would be seen consistently across all LST's, which is not what is seen here.

It is difficult to trace back the causes of these non-redundancies to individual antennas from

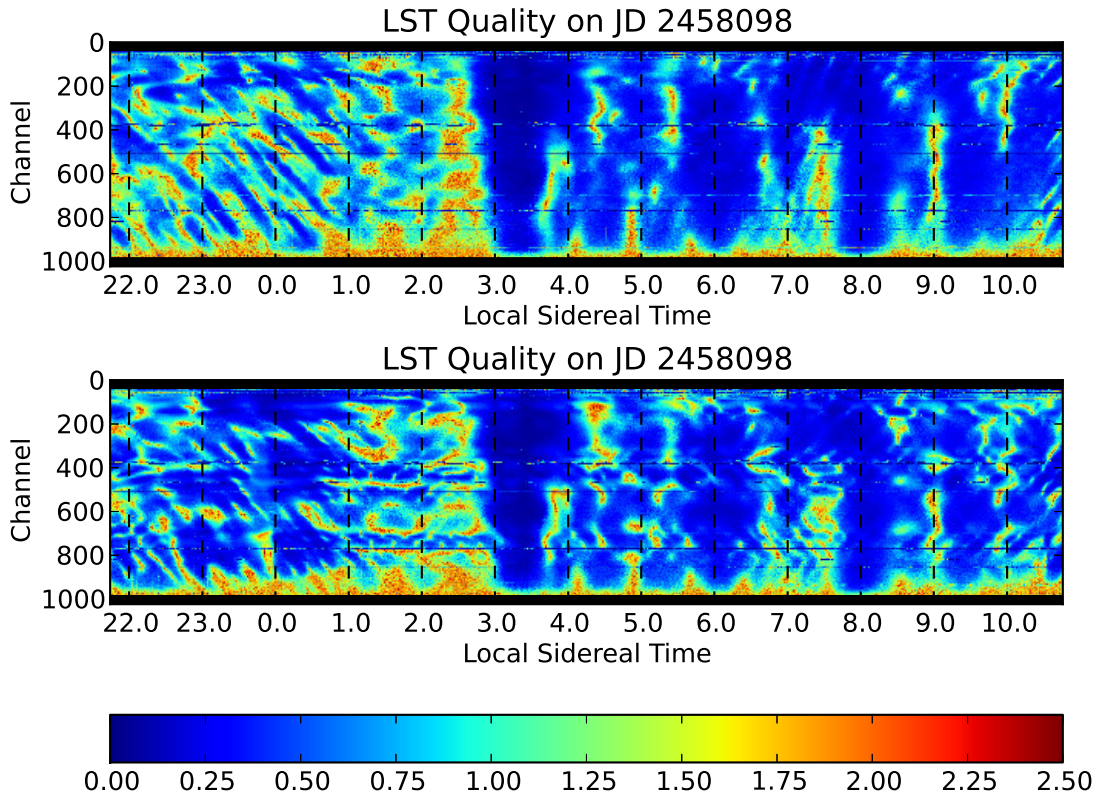

Figure 4.4: Redundancy for East-West Triads X (top) and Y (bottom) polarisations for JD 2458114. The colour scale is in units of standard deviation of the phases, at a certain sidereal time and frequency channel.

the closure phases alone, which constitutes a fairly intractable inverse problem, even more intractable than the same problem with visibilities due to being a further stage removed. However as shown here the closure phases provide an excellent and stable statistic, in combination with modelling, quantifying direction-dependent non-redundancies and other non-redundancy terms such as closure errors caused by correlated noise terms between antennas.

Modelling a HERA-like array with an Airy disc beam profile with different beam errors, and calculating redundancy against a sky model built on the GLEAM catalogue, confirm that beam deviations between antennas is a very strong driver of non-redundancy. In addition to this, correlated systematics between antennas at a level of 1% of the signal is enough to greatly degrade redundancy in our simplified model.

It is known that HERA suffers from mutual coupling between antennas on the order of 1 part in 10^4 which results in a complex deviation of antenna beams from each other. The new Vivaldi feeds are expected to make this problem worse. Fortunately, feed locations are likely to be more carefully controlled in future with the addition of the new feeds, in concert with a laser positioning system for accurate feed positioning.

Overall, by using the closure phases of a group of redundant triads, the ability is demonstrated to measure non-redundancy across a redundantly arranged array such as HERA, CHIME or HIRAX as a function of LST and Frequency. Modelling of the array indicates that the non-redundancy between antenna beams is very complex with pointing errors being one of the largest drivers of non-redundancy, if not the largest.

A method for mapping the redundancy of the HERA telescope using the closure phases is shown. The work by Fagnoni (Fagnoni et al. 2019), in parallel with these results, suggests that there is a problem with divergence between direction dependent antenna responses across the HERA array. Closure phases provide a powerful diagnostic for measuring the redundancy of an array such as HERA, where redundant baselines are one of the key tenets of the design engineering.

This diagnostic makes use of the independence from antenna specific gains; this is not available in visibilities, which are vulnerable to calibration errors. It is important to note that the modelling allows major contributors to the non-redundancy statistic to be measured, but is unable to stratify the individual contributions to the overall non-redundancy statistic. As stated, this constitutes a quite, but not entirely, intractable inverse linear problem, with a degenerate solution space. Additionally, it is not fully capable of narrowing problems down to individual antennas, as not all antennas in an array are guaranteed to be in two or more identical triads, such as antennas on the edge of the array. This is nevertheless a powerful technique for diagnosing the overall health of current and future interferometric arrays which rely on redundancy as a major design characteristic.

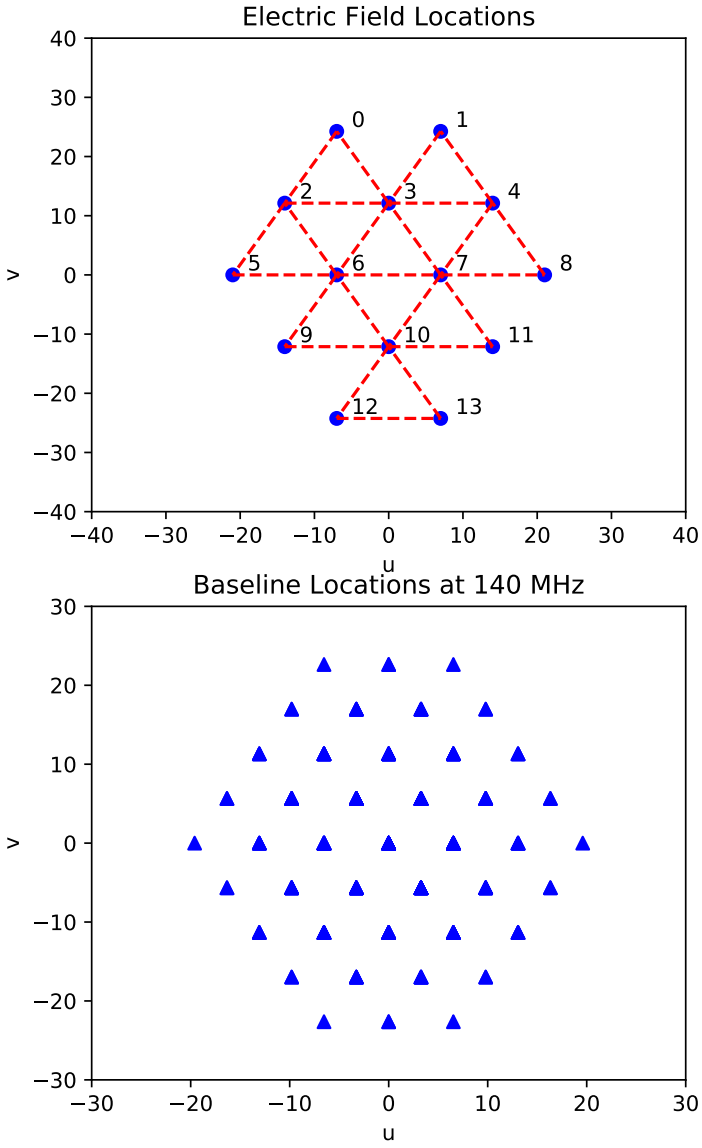

Figure 4.5: Simulated Antenna locations and resultant baseline locations. The overlays in the top figure indicate the equilateral triads chosen. Only North facing equilateral triads were chosen for this simulation.

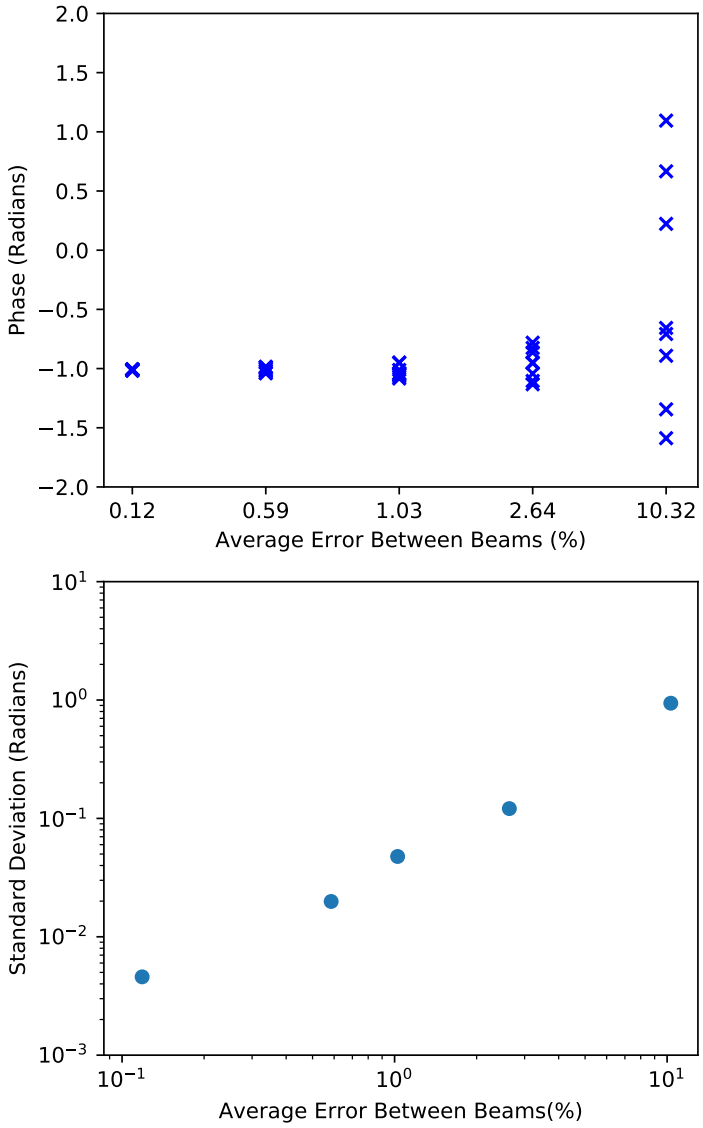

Figure 4.6: A simulated measure of how the closure phases diverge as a function of the beam error. The beams of the instrument are randomly varied and the average beam error is measured and shown, and from this the closures are measured.

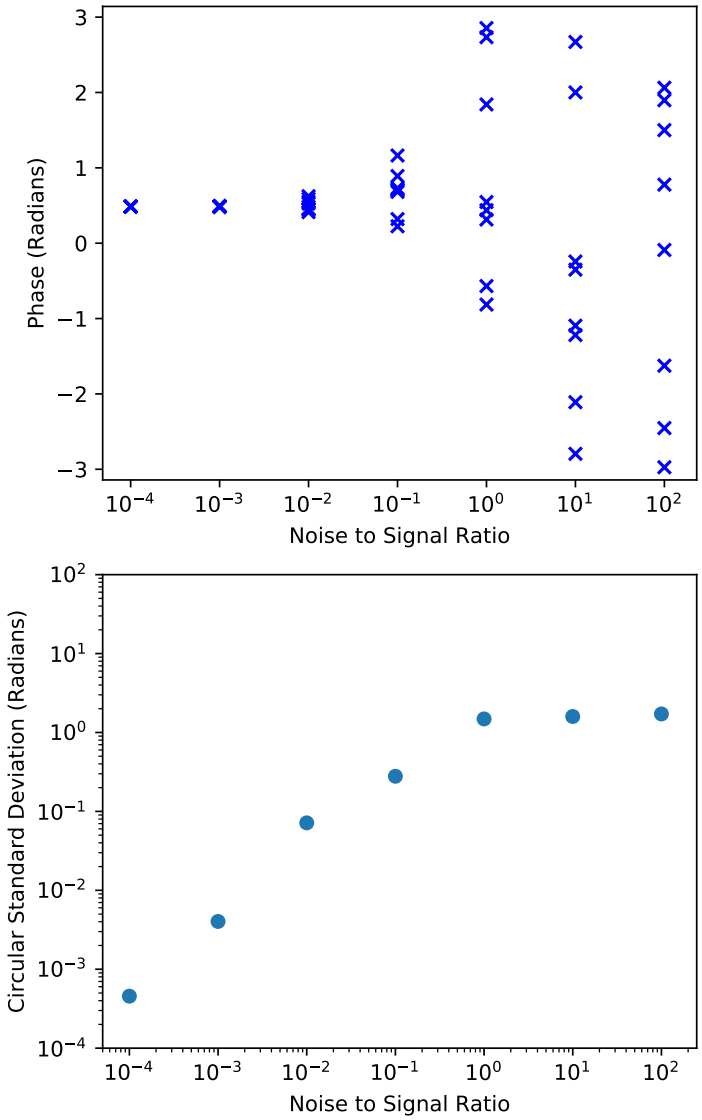

Figure 4.7: A simulated measure of how the closure phases diverge as a function of additive correlated baseline noise. This could be a signal such as the HI signal, or a correlated term such as a standing wave.

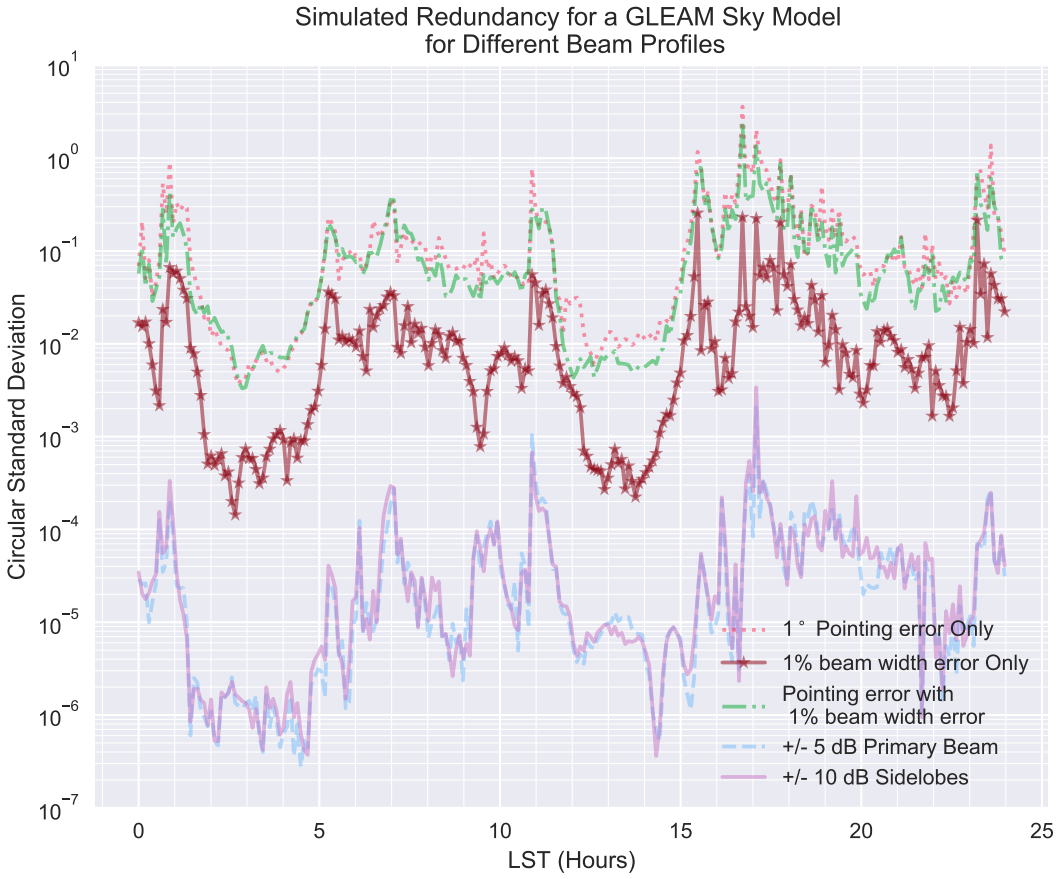

Figure 4.8: The redundancy of the array shown in Figure 4.5 vs Local Sidereal Time, for Equilateral 14m Triads, for a sky model consisting of GLEAM sources with a flux above 8 Jy. The Fornax A window at LST 3h 25m has a low divergence between closure phases which is expected for a strong source dominating the field of view. The beams of each antenna are an airy disc function, and modified to incorporate pointing error, beam width errors, and primary and sidelobe gain errors. Pointing error is by far the strongest driver of non-redundancy in this model. The standard deviation of the standard deviations plotted is roughly 0.27σ, with 0.27 being a standard calculated weight based on the number of samples.

Chapter 5

A limit on the 21-cm Epoch of Reionisation Signal using the Bispectrum

5.1 Introduction

The results in this chapter were achieved in collaboration with colleagues from Cambridge and the USA. This work and methods described herein, has been accepted to Physical Review Letters D (Thyagarajan et al. 2020), and the methodology contained within this chapter, specifically the data processing pipeline, has also been used in a modeling effort for HERA (Carilli et al. 2020). All analysis presented in this chapter has nevertheless been replicated and implemented directly by myself.

In the past few years, several telescopes have been built with the goal of discovering the redshifted 21-cm spectral line from neutral hydrogen. The 21-cm line is due to a quantum spin-flip transition (Feynman et al. 1965). It acts as a probe of the cosmological timeline by virtue of it being redshifted. It is possible, by measuring the red-shifted 21cm line, to quantify the Epoch of Reionisation (EoR), during which neutral hydrogen was ionised by the first stellar objects, such as star-forming regions and active galactic nuclei (AGN's) (Furlanetto 2016).

Numerical simulations of the cosmological evolution of the universe have predicted the fluctuation in the 21-cm signal would begin at a redshift value of $z = 25$ and be well underway by $z = 20$. In this Epoch, the hydrogen gas in the universe is undergoing heating from a combination of Lyman-α and X-Ray sources, and this moves into the EoR at a redshift value of $z = 15$ (Furlanetto 2016). The EoR progresses until full ionisation of the neutral hydrogen

in the universe is completed at a redshift of approximately $z = 8$.

The instrumental difficulties in directly detecting the brightness fluctuations in the 21-cm signal background are, however, immense. The 21-cm signal is faint in its rest frame (compared to other spectral lines of hydrogen), with this being compounded by the exceptionally bright radio foregrounds that exist at the redshifted frequency of the 21cm line during the EoR. Simulations such as those discussed in Furlanetto (2016) have shown that the ratio in the power of the foregrounds to the power of the 21-cm signal is roughly 5 orders of magnitude. Looking for this 'needle in the haystack' requires sophisticated data analysis techniques in combination with a suitably designed radio telescope.

Arrays that are being designed around the search for the 21-cm line in the EoR include the Hydrogen Epoch of Re-ionisation Array (HERA) (DeBoer et al. 2017), The Hydrogen Intensity and Real-Time Analysis eXperiment (HIRAX) (Newburgh et al. 2016), the Canadian Hydrogen Intensity Mapping Experiment (CHIME) (Bandura et al. 2014), as well as the planned Square Kilometre Array (SKA) (Dewdney et al. 2009a). The analysis and limits derived in this chapter are done with the HERA array in South Africa, in concert with the work in Chapter 4.

In this Chapter, we will use the Bispectrum of the closure phase to do an initial data analysis of the Epoch of Reionisation using the HERA telescope. These results were derived using the Delay Spectrum (Parsons et al. 2012) technique, on two frequency bands centred on redshifts $z = 10.45$ and $z = 7.7$. The delay spectrum technique is discussed in Section 5.2, and the data analysis pipeline in Section 5.3. The conversion from closure phase units to cosmological units is discussed in Section 5.4, with results in Section 5.5, and a discussion about the ramifications for future 21-cm analysis in Section 5.6.

5.2 Delay Spectrum

The closure phase is constructed using the mathematics discussed in Chapters 1 and 4. Using this same method of selecting triads in a set that are of equal size and orientation, it is possible to detect the phase fluctuations from the EoR, as discussed in Thyagarajan et al. (2018). The fluctuations from the EoR will vary as a function of red-shift over the line of sight, as well as orthogonal to the line of sight. This is a result of the bubbles of ionised hydrogen that formed during the Epoch of Reionisation, following ionisation by the hard X-Ray and Ultraviolet background produced by the first stellar structures (Furlanetto 2016).

These fluctuations in frequency can be detected by a low frequency radio interferometer such as HERA (DeBoer et al. 2017). HERA's redundant layout allows many redundant baselines, and consequently many redundant closure triads, which allows different cosmic scale structures to be sampled. Shorter baselines have a fringe interference pattern broader than those of longer

baselines. These fringes, a measurement of the spatial coherence function as discussed in Chapter 1, additively added together from all baselines, is how an interferometer forms an image by the van-Cittert Zernike theorem (Born and Wolf 1999).

These fluctuations can be measured by using a delay spectrum, as discussed in Parsons et al. (2012), which takes the Fourier transform of the spectrum of visibilities constituting a set of redundant baselines. First define a visibility, defined in Equation 1.16, spectrum for a single baseline $V(\bar{u}, \bar{v}, \eta)$, then take the Fourier transform:

$$D_\nabla(\tau, \eta) = \int V(\bar{u}, \bar{v}, \eta) e^{-2\pi i \eta \tau} d\eta \tag{5.1}$$

where τ represents delay in seconds, η represents frequency, and $\bar{u}$ and $\bar{v}$ represent a baseline of a redundant set (that have the same u/v values). In this analysis w has been deliberately neglected as the HERA array is flat, and therefore the w term discussed in previous chapters is negligible. Starting from the van-Cittert Zernike theorem in Equation 1.16, Equation 5.1 can be expanded to:

$$D_\nabla(\tau, \eta) = \iiint I(l, m, \eta) e^{-2\pi i(ul+vm)} e^{-2\pi i \eta \tau} dl\, dm\, d\eta \tag{5.2}$$

Upon defining geometric delay as $\tau_g = \frac{\eta}{c}(b_x l + b_y m)$, where b_x and b_y are the x and y components of the baseline vectors in meters in some coordinate frame of reference. This equation can now be reduced to:

$$D_\nabla(\tau, \eta) = \iiint I(l, m, \eta) e^{-2\pi i \eta(\tau_g - \tau)} dl\, dm\, d\eta \tag{5.3}$$

5.2.1 Closure Delay Spectrum

The delay spectrum approach can also be done using a set of redundant closure phase triads, as discussed in Thyagarajan et al. (2018). The major advantage to using the closure delay spectrum over the visibility delay spectrum is that closure phases are naturally immune to direction-independent antenna terms, such as complex gains in the electronics chain. Working with visibilities, these terms have to be solved for and corrected in the form of calibration, to end up with a coherent image or statistical quantity. Closure phases, apart from correction of direction-dependent effects and closure errors discussed in Chapter 4, do not require calibration. This is therefore an alternative and attractive method to extracting the EoR signature from a suitably designed low frequency interferometer.

To quantify this, let $C_\nabla(\eta)$ denote a closure phase spectrum from a triad that is in a set of redundant triads. From this we can define the Closure Delay Spectrum:

$$D_\nabla(\tau) = \int C_\nabla e^{-2\pi i \eta \tau} d\eta \tag{5.4}$$

By taking the averages of the closure cross-power spectral density over redundant triads at the same time, the thermal noise can be reduced. The noise that exists on different baselines is completely independent even in the case of an antenna being shared between two baselines, and by extension, two triads; this is shown in Chapter 1, and demonstrated rigorously in Appendix A.

In the data analysis considered in this chapter we have not averaged over multiple triad groups for the reasons that we consider, as will be shown, the data is systematics limited, and that the necessary number of averages required to reduce the thermal noise to the level of the EoR signal is very high. However with the results from Appendix A in mind, the analysis can be extended to include all triads which do not share a baseline. This represents a combinatorics problem which is an avenue for future work, and will not be considered further here.

5.3 Data Analysis Pipeline

The data from HERA consists of raw, uncalibrated visibility data from all antennas. Each visibility constitutes a 'session', with a visibility in each session being averaged for 10.7s to reduce the thermal Gaussian noise from each measurement. Assuming the noise is described by a Gaussian distribution, and each noise measurement is considered statistically independent, the overall noise reduces as $\sqrt{N}$, with N being the number of measurements. An overview of the entire data analysis pipeline is shown in Figure 5.1.

5.3.1 Closure Generation

The stored visibility data is then processed to form closure phases for different redundant sets of triads. The redundant sets chosen were as follows:

- Equilateral 14m Triads
- Equilateral 28m Triads
- Linear East-West Triads - Consists of two 14m baselines and a 28m baseline, oriented East to West.
- Linear P30 Triads - Consists of two 14m baselines and a 28m baseline, oriented at 300 degrees from the North.
- Linear M30 Triads - Consists of two 14m baselines and a 28m baseline, oriented at 60 degrees to the North.

A plot of the HERA array layout is shown in Figure 4.1. The closure phases are formed into a data cube of shape [Time, Day, Triad, Channel], with a seperate data cube for each set of redundant triads. No calibration is applied because, as discussed, closure phases are

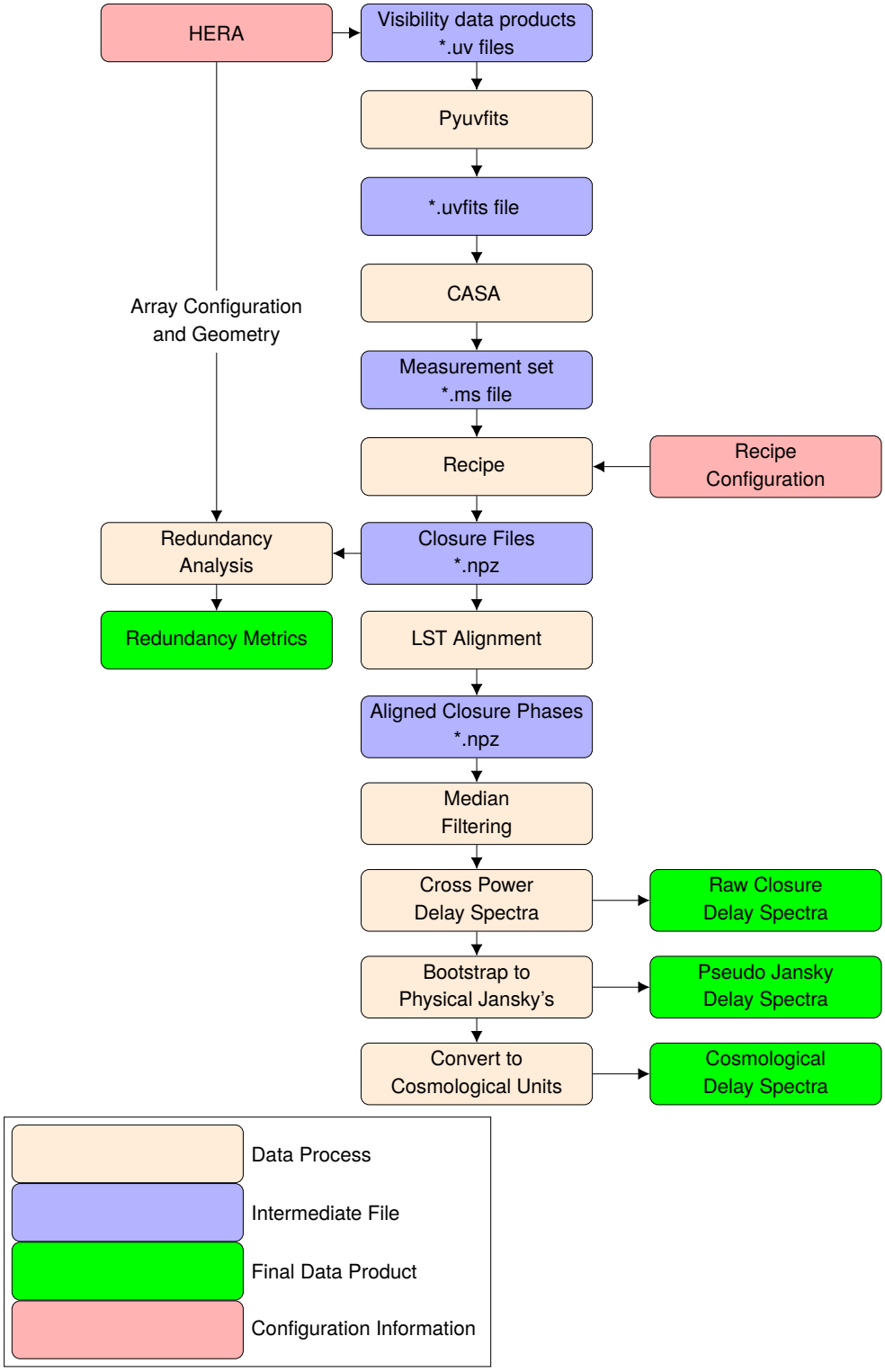

Figure 5.1: Data pipeline for the bispectrum analysis, including the redundancy analysis covered in Chapter 4

immune to antenna-dependent direction-independent gain terms. The conversion of the miriad files to Measurement Set files and the extraction of the closure phases was performed using the Recipe framework (Nikolic et al. 2018). This allows minimal recomputation, where previously processed and generated files are cached to allow quick retrieval when re-running the data processing pipeline, or running with minor modifications. This vastly decreased the amount of time waiting for time intensive file conversions to complete.

5.3.2 Local Sidereal Time Alignment

After the formation of the closure phases it is necessary to align them so that the field of interest is aligned in the Time axis. The reason this is necessary is because of the natural drift between the solar day and the sidereal day. This drift is on the order of roughly 4.5 minutes a day. HERA observes between the times of 8pm and 6am Greenwich Mean Time (GMT), and the fixed observation start and end times in the solar period together with the LST drift means that there are roughly 22 sessions difference between a LST on one day and a LST on the next day.

For the present analysis, alignment was performed for all days in the IDR2 dataset (Dillon 2018), between JD 2458098 and JD 2458118. The alignment utility is written in Python, and works by lining up closure sessions to within 10.7s for each LST [a]. It has been shown that variation within each session should be negligible, and that averaging can even be made over sessions for up to 1-2 minutes to further reduce the Gaussian thermal noise (Carilli et al. 2018).

The field that was aligned and extracted for this analysis was the Fornax A field. The dataset spans between LST 02:30 and 04:30. This field was chosen because of the strong point source dominating the primary beam of HERA, which suppressed the severe non-redundancies in the telescope, discussed in Chapter 4. Outside of this field, the redundancy is of such a poor nature that it degrades the dynamic range in the consequent delay spectra.

5.3.3 Data Reduction and Delay Spectrum Formation

After alignment of each redundant set of triads in LST for the Fornax A field, a delay spectrum was formed within each redundant set. At this point, the data is in the form of a 4-dimensional 'cube' of data of shape [Time, Day, Triad, Channel], with the Time axis now referring to Local Sidereal Time following alignment.

A median filter was then applied across the two halves of the Day axis. By taking the median filter across this axis, Radio Frequency Interference (RFI) can be filtered out. Both narrowband and broadband RFI corrupts the spectrum between the 100 and 200 MHz at which HERA operates. Our analysis does not utilise flagging of RFI corrupted channels, because of

[a] Source code for the alignment software is available alongside this thesis.

the necessity to have continuous tracts of spectrum to perform the delay spectrum. It may be possible to interpolate over channels but this was not thought feasible due to the very slight and high frequency oscillations of the 21-cm signal as a function of frequency. Interpolation was not explored further, and instead the segments of spectrum were selected with the least amount of corrupting RFI, based on inspection of the data. After the median filtering, the shape of the data cube is [Time, 2, Triad, Channel].

Two sub-sections of the HERA band were selected from which to produce delay spectra. One section was from channels channels 120 to 380 of the correlator (the 'Low Band'), and the other from channels 515 to 775 (the 'Upper Band'). The Low Band corresponds to frequencies 111 MHz to 137 MHz, and the Upper Band a frequency range from 150 MHz to 176 MHz. In the centre of the Low Band this corresponds to a redshift value of z=10.45, and the Upper Band z=7.7. In the Low Band this should correspond to a time during the main stage of reionisation, and the Upper Band towards late reionisation and the end of the Epoch of Reionisation (Furlanetto 2016).

For these two bands, a Cross Power Spectral Density (CSD) measurement was performed between one half of the median filtered days and the other half. Let $C_{\nabla,a}(t,\nabla,\eta)$ and $C_{\nabla,b}(t,\nabla,\eta)$ denote these two median filtered quantities. From this the CSD can be defined:

$$\bar{C}_{\nabla,ab} = C_{\nabla,a}(t,\nabla,\eta) * C_{\nabla,b}(t,\nabla,\eta) \tag{5.5}$$

where $*$ denotes a convolution of all triads and times against the same triad and time in each set of median filtered closure phases. Cross-correlation was not performed across different triads because of the risk of degradation of the HI signal due to the divergence in response in a supposedly redundant set (discussed in the previous Chapter). Each closure triad, within a set of the same type, has a different response, likely derived from pointing errors and other direction dependent effects. This could cause the HI signal to decorrelate if cross-correlating across triads. Using the Wiener-Khinchin theorem, Equation 5.5 becomes a multiplication of the Fourier transforms of $C_{\nabla,a}(t,\nabla,\eta)$ and $C_{\nabla,b}(t,\nabla,\eta)$:

$$\bar{C}_{\nabla,ab} = \left[\int C_{\nabla,a} e^{-2\pi i\eta\tau} d\eta\right] \cdot \left[\int C_{\nabla,b} e^{-2\pi i\eta\tau} d\eta\right] \tag{5.6}$$

$$\bar{D}_{\nabla,ab} = D_{\nabla,a}(\tau) \cdot D_{\nabla,b}(\tau) \tag{5.7}$$

The reason behind using the cross-spectral density is that all terms between the two Closure spectra that are uncorrelated have an expectation value of zero, and should average out. Correlated terms, such as the HI signal and correlated systematics, have a non-zero expectation value, and will represent any signal after averaging that is not a thermal-like noise signal. To achieve this, the data cube of delay spectra is then averaged over time (not delay), and over

triads within a redundant set, so as to reduce thermal noise, which is assumed to be independent and Gaussian. This produces raw delay spectra with varying dynamic ranges, as seen in Figures 5.2, 5.3, 5.4, 5.5.

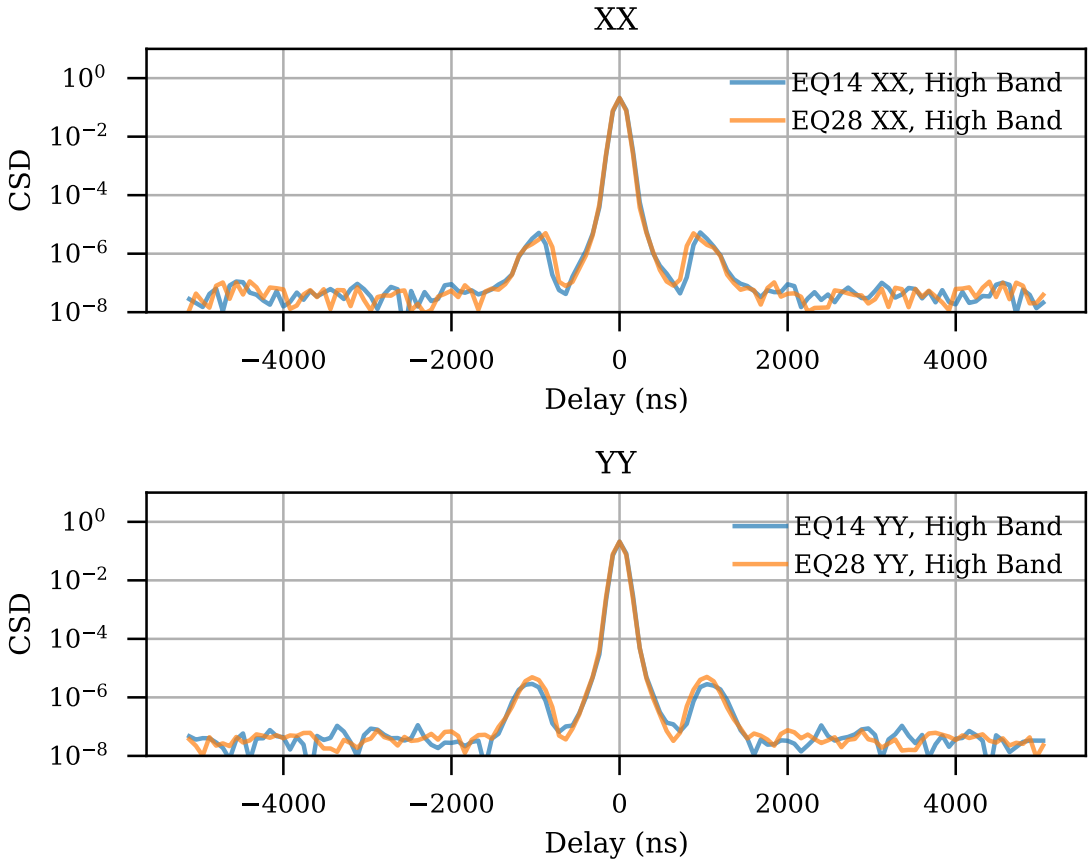

Figure 5.2: Delay Spectra of Equilateral Triads in the Upper Band

5.4 Physical Units Interpretation

The absolute closure delay-spectra are presented in arbitrary units within closure space bounded on the domain $[0, 1]$, with this domain restriction a consequence of the phasor being defined on the unit circle. It is necessary to relate the closure spectra shown in Figures 5.2, 5.3, 5.4 and 5.5 to physical units, first by an approximate conversion to flux, and then onto physical cosmological units in terms of $mK^2h^{-3}Mpc^3$, which is the de-facto standard in literature (such as Parsons et al. (2012)).

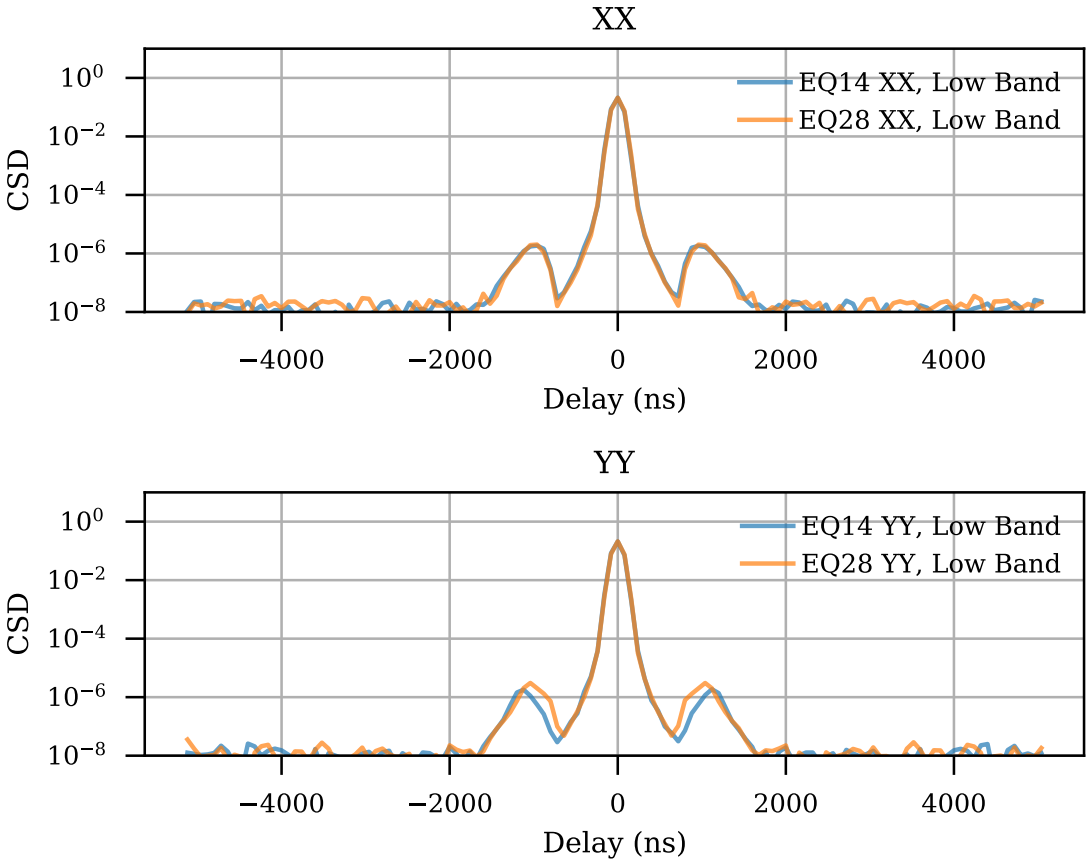

Figure 5.3: Delay Spectra of Equilateral Triads in the Low Band

5.4.1 Bootstrapping to Physical Jansky's

We now limit our attention to the Equilateral triad classes consisting of baselines of 14m and 28m in length, respectively. The closure phase is defined in Equation 1.37 as the angular component of the complex vector formed by the multiplication of three visibilities composing a closed geometric triad. Thus, although the angular component provides our cosmological probes in terms of $k_{||}$ and $k_{\perp}$, it is naturally defined as a multiplication of three complex numbers representing, arbitrarily at this point, the flux of the spatial coherence mode measured by each baseline in the triad.

To convert the delay spectra presented in Figures 5.2 and Figures 5.3 to cosmological unit, they must be first converted to this flux component. To this end we use a set of reference calibrated visibilities, which have been generated by the HERA collaboration using OMNICAL, which is a redundant calibration of the (supposedly) redundant baselines used in each triad. We can then take a scaling value for our delay spectra by adding in inverse quadrature the flux

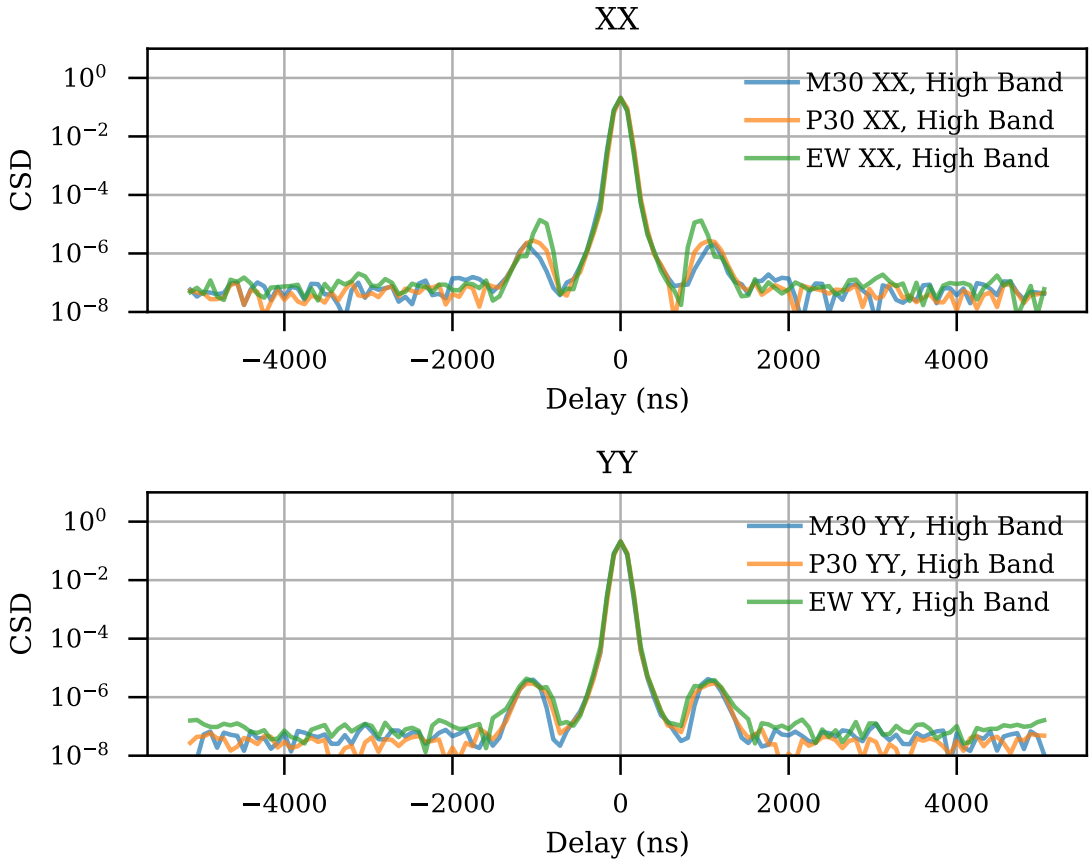

Figure 5.4: Delay Spectra of Linear Triads in the Upper Band

values from each baseline:

$$A_\nabla = \sqrt{V_{ab}V_{bc}V_{ca}}^{\frac{1}{3}} \tag{5.8}$$

$$\bar{A}_\nabla = A_\nabla^2 = V^2 \tag{5.9}$$

Use of the the values provided from the redundant calibration of the HERA visibilities for the IDR2 dataset used here makes it possible to put the delay spectra back into a physical unit. We assume that the variation of the Jansky value within our chosen band does not vary by a large amount, which has been empirically validated. However, the closure spectra essentially form a surjective, many-to-one linear map of the form:

$$M_\nabla : V \mapsto C_\nabla \tag{5.10}$$

with M_∇ representing our map from visibilities, V, to closure C_∇. Thus there is not a unique reverse to M_∇ allowing a direct mapping back to the sky variables. However, multiple sources,

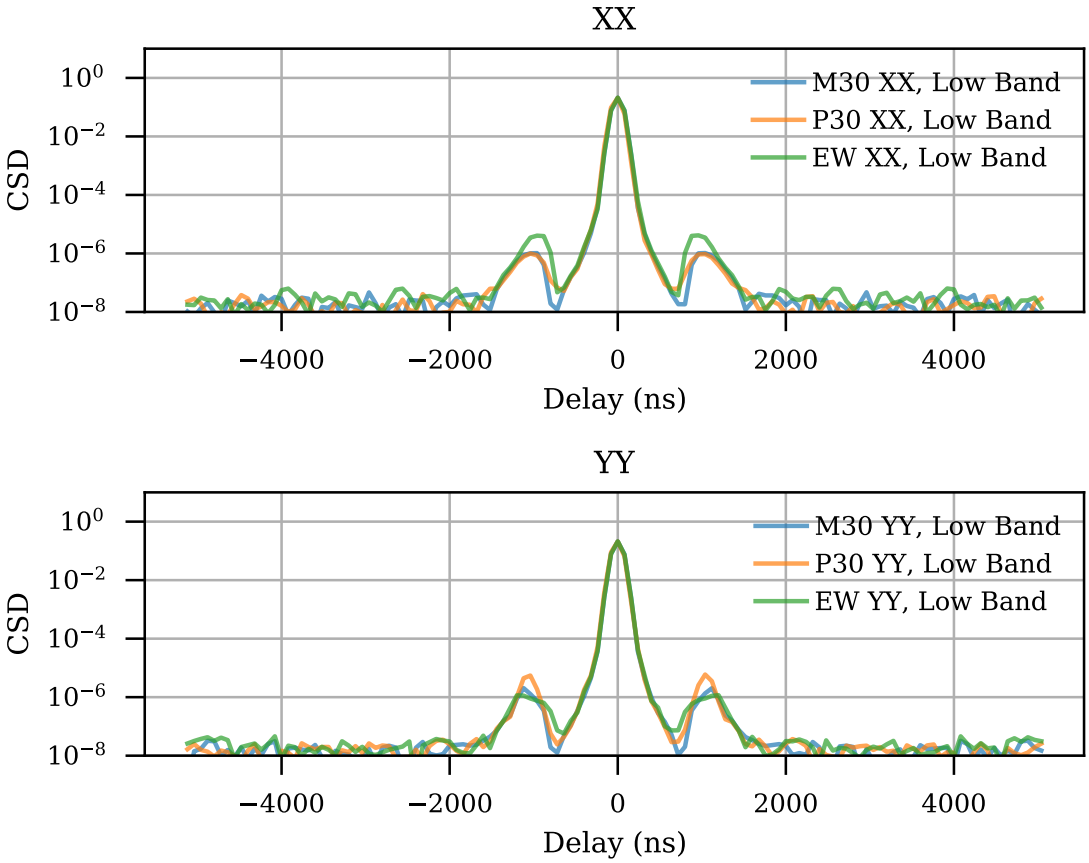

Figure 5.5: Delay Spectra of Linear Triads in the Low Band

such as Blackburn et al. (2019) and Kulkarni (1989), note that whilst going from closure phases to visibilities is an inverse problem, the closure phases are direct observables of the sky, with each representing a unique degree of freedom. Thus our approximation can be considered physically informed, and able to provide an astrophysical interpretation, even in the absence of an exact map comprising the inverse of M_∇. We can now multiply our delay spectrum by $\bar{A}_\nabla$ to gain a delay spectrum in units of $JyHz$, which is shown in Figure 5.6. The real component of the delay spectrum in Figure 5.6 is shown in Figure 5.7.

5.4.2 Conversion to Cosmological Units

The physical quantities shown in Figures 5.6 and 5.7 represent a flux in the frame of reference of the telescope for each $k_{||}$ bin. The final step is to convert this flux to a comoving cosmological scale taking into account the cosmological parameters of the ΛCDM model, as well as the properties of the telescope. A method of conversion from the telescope flux units to our

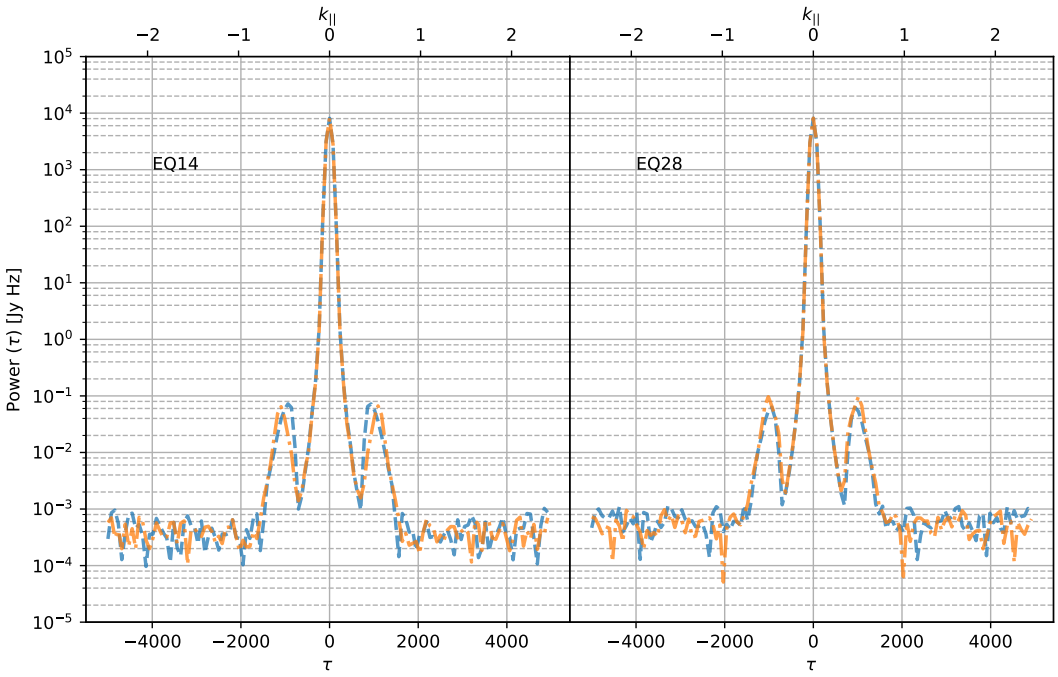

Figure 5.6: The absolute delay spectrum after converting to a power spectrum in Jansky's. There is not a direct mapping between the closure phase and the physical sky, but this allows the power spectrum to be interpreted in a way which harmonises with existing literature.

cosmological units is of the form:

$$P(k_{||}) = \frac{X^2 Y}{\Omega B \nu_{21}^2} \left(\frac{\lambda^2}{2k_B} \right)^2 D_{\nabla,ab} \tag{5.11}$$

where $X = c(1+z)^2 H_0 E(z)$, $Y = D(z)$, with c as the speed of light, H_0 is the hubble parameter, and $E(z) = [\Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda]^{\frac{1}{2}}$ which is a direct consequence of the Friedmann Equations, representing an exact solution to the Einstein field equations in a Friedman-Lemaitre-Robertson-Walker (FLRW) metric. Here, $D(z)$ represents the comoving radial distance, Ω the angular area of the primary beam of the telescope, B the bandwidth represented by the delay transform in Hz, ν_{21} the rest-frame frequency of the 21-cm spin-flip transition of the hydrogen atom, λ the wavelength of the centre of the observation band; k_b as the boltzmann constant.

The cosmological scalar conversion in Equation 5.11 is calculated using the HERA PSPEC routines. The cosmological parameters used to make this conversion are of those given in the the Planck 2015 results (Planck Collaboration et al. 2016); a summary of these is given in Table 5.1. The angular area of the primary beam is approximated using a Gaussian beam with a full Width at half maximum of 6°. The resulting scalar is multiplied into the delay spectrum

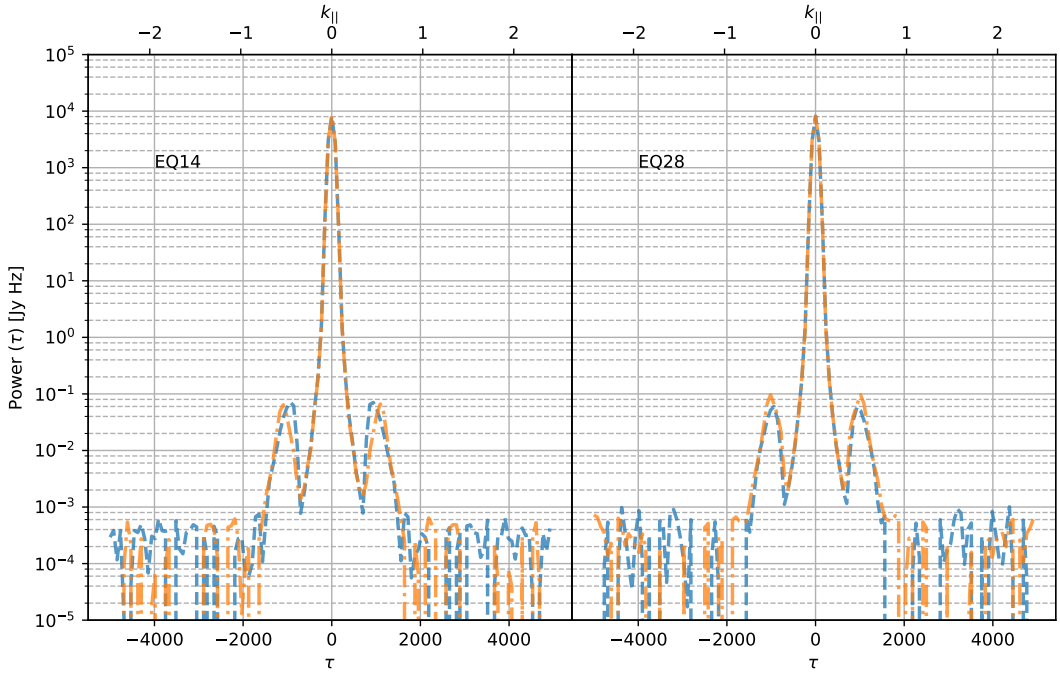

Figure 5.7: The real delay spectrum after converting to a power spectrum in Jansky's. There is not a direct mapping between the closure phase and the physical sky, but this allows the power spectrum to be interpreted in a way which harmonises with existing literature. The zero crossings outside of the foreground wedge and systematics indicate that we are still noise dominated and that there is no definitive detection of a signal that is not the foregrounds or instrumental systematics.

Parameter	Symbol	Value
Dark Energy Density	Ω_Λ	0.6911
Matter Density	Ω_m	0.3089
Spatial Curvature	Ω_k	0.0008
Hubble Constant	H_0	67.66 $kms^{-1}Mpc^{-1}$

Table 5.1: Cosmological parameters used in the conversion of the delay spectra to cosmological units. These parameters are from the Planck 2015 results (Planck Collaboration et al. 2016)

in Figures 5.6 and 5.7 along with a conversion from Janskys to milliKelvin, to produce the cosmological delay spectra seen in Figures 5.8 and 5.9.

5.5 Results

A set of delay spectra using the closure method are now presented in cosmological units. The dynamic range achieved is on the order of 8 units of magnitude, which is broadly consistent

with the other delay spectra methods. There is no detection of the 21-cm signal using this method with the current dataset.

One major conclusion is that we are limited by the systematics shown in the delay spectra, especially around a τ value of $1\mu s$, which is hypothesised to be from reflections in the cables running from the antenna feeds due to impedance mismatch effects. This systematic bump completely covers the delay bins in $k_{||}$ which are most sensitive to the spin-flip temperature, representing the EoR.

In Figures 5.7 and 5.9 we notice that outside of the systematics limited region, there are multiple zero-crossings of the real component of the delay spectrum. This indicates that, systematics notwithstanding, we are ultimately limited by the thermal noise. A detection should show up as a defined signal within the real component which does not have a zero-crossing.

As noted in Thyagarajan et al. (2018), roughly 10 orders of magnitude in dynamic range is required to begin to uncover the frequency-dependent oscillations in the spin temperature. Here we are at least two orders of magnitude away from that, assuming that there is no decoherence of the 21-cm signal due to instrumental effects. In reality, considering cosmological simulations of the 21cm line in Mesinger et al. (2011), and noted by others [Thyagarajan, Private Communication] the cosmological signal lies somewhere in the region of 10^1 to $10^4 mK^2 h^{-3} Mpc^3$ in the window analysed here, with it being weaker at lower redshifts. Thus we remain significantly thermal noise dominated in the absence of systematics, and even the absence of systematics under the current noise floor is unlikely.

5.6 Discussion

A functioning data analysis pipeline is presented for a delay spectrum method utilising the Bispectrum. Initial limits have been presented with the Fornax A field, which was chosen due to the excellent redundancy characteristics of the telescope, as seen in Chapter 4.

The resultant delay spectra are fully capable of probing the 21-cm spin flip temperature of the Epoch of Reionisation, as laid out in Thyagarajan et al. (2018). The present analysis has concentrated on a field with particularly strong foreground emission due to the presence of Fornax A. This has the benefit of holding the redundancy of the telescope at a good level relative to other fields, but is not an ideal observation strategy. This is because the high foreground emission makes it harder to achieve the high dynamic range required to measure the HI signal. In reality, it will be necessary to average iteratively over many fields, excluding Fornax A and any data where it exists within the main beam of the telescope, because of its propensity to dominate the overall telescope response.

Several distinct issues arise at once in analysing this data. Redundancy has been considered

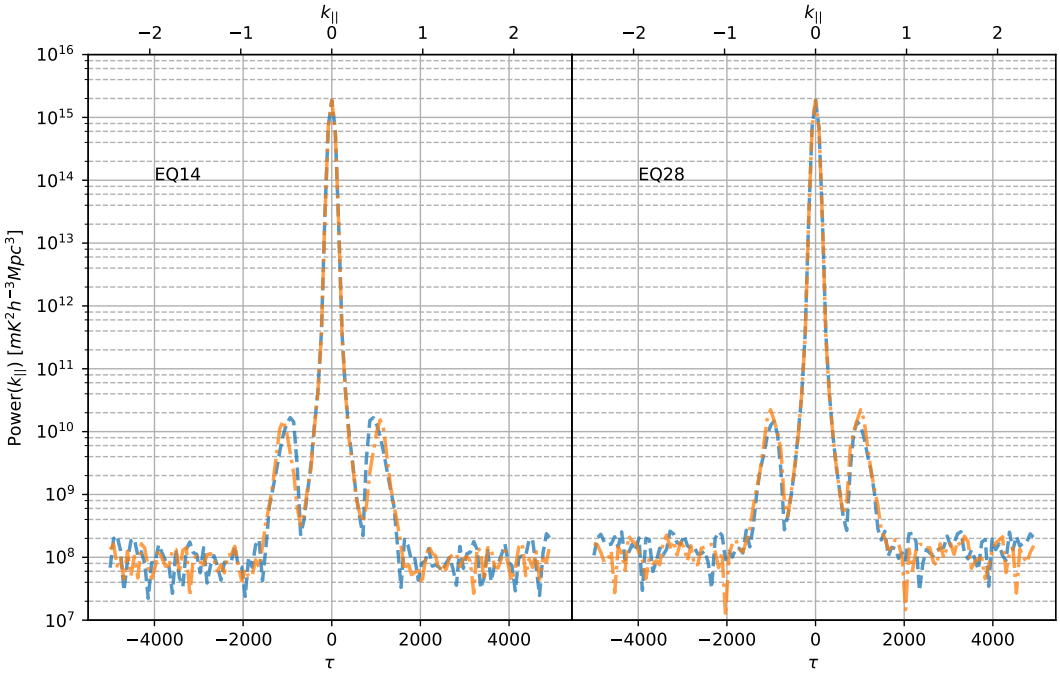

Figure 5.8: The absolute delay spectrum after converting from the power spectrum in Figure 5.6 to a power spectrum in cosmological units of $mK^2h^{-3}Mpc^3$, by the conversion shown in Equation 5.11. There is not a direct mapping between the closure phase and the physical sky, but this allows the power spectrum to be interpreted in a way which harmonises with existing literature.

in detail in Chapter 4, and we add only that non-redundancy to such a great degree as has been demonstrated voids many assumptions that go into the delay spectrum pipelines of the bispectrum method and even the use of visibilities. For the bispectrum method, each closure phase is no longer drawn from the same probability distribution, so that averaging over the delay spectra is a dubious procedure, with the potential for degradation or perhaps even the overall loss of the faint 21-cm signal.

The non-redundancy is of a direction dependent nature, and requires further research to correct for, and to understand the effects on the weak signal of the EoR. Other systematics clearly exist in this data at a significant level, with the 'bumps' either side of the main foreground peak in the delay spectra likely to be due to some reflection effect within the cables. The HERA telescope will be upgraded over the next year with an optical fiber system, which will significantly shorten the cable to the point that any reflection systematic will be moved away from the EoR window in the delay spectra.

Our currently noise limited measurements outside of the systematics limited k-bins will

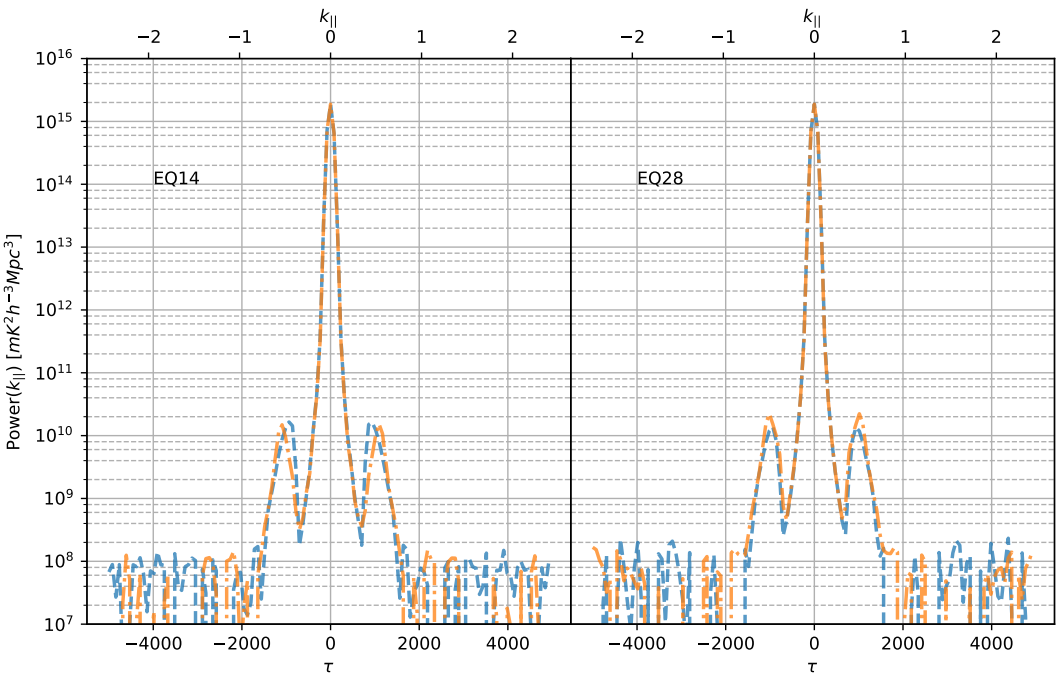

Figure 5.9: The real delay spectrum after converting from the power spectrum in Figure 5.7 to a power spectrum in cosmological units of $mK^2h^{-3}Mpc^3$, by the conversion shown in Equation 5.11. The zero crossings outside of the foreground wedge and systematics indicate that we are still noise dominated and that there is no definitive detection of a signal that is not the foregrounds or instrumental systematics.

require more integration time over a greater number of fields, and a greater number of frequency bands. Additionally, more triad classes must be used to continue to drive down the thermal noise. As noted in Appendix A, only triads that share a baseline should be covariant, so there is great scope in continuing to drive down the thermal noise.

By noticing the systematic peak in the delay spectra rises above the thermal noise of all k-bins, we conjecture that other systematics exist below the thermal noise which have not yet been discovered. Knowledge of the instrument, and of its effects on the data, is of paramount importance, with the caveat that the effects of systematics and non-redundancies on very small signals such as the EoR should be explored in order to determine what level of signal loss can be expected.

The bispectrum delay spectrum using the closure phase is an excellent tool in the arsenal of astronomers seeking to uncover and probe the EoR. Its immunity to direction-independent effects means that calibration can be forgone. It is still vulnerable, however, to unchecked non-redundancies from direction-dependent effects, and further research is required on constraining

and correcting for these. A limit has been presented here which is currently limited by both systematics and noise. Additional integration over a greater number of nights, and fields, and triad classes, should reduce the thermal noise floor, in parallel with instrumental improvements to reduce systematic errors.

CHAPTER 6

HIGH PERFORMANCE WIDE-FIELD IMAGING

6.1 Introduction

Radio astronomy has always been at the forefront of computing techniques, both in software and hardware. The ability to construct maps accurately and quickly has to some extent always been limited by modern compute hardware, and radio astronomy has consistently been one of the main drivers of new digital signal processing hardware and software, as well as stimulating new mathematical methods for aperture synthesis techniques. In the modern vernacular, it has been and likely always will be a discipline concerned with "big data".

To emphasise the parallel histories of radio interferometry, with its usage of aperture synthesis mapping techniques, and that of high performance computing, let us look at the development of electronic computers in the 1950's. The computer laboratory of the University of Cambridge pioneered in the development of electronic computers, especially with Edsac I (1949-1957), Edsac II (1957-1965), and Titan (1963-1973). These were all the most powerful computers available in the world during their operation. The first aperture synthesis mapping done at the Cavendish Laboratory was performed on these computers, taking large amounts of computer time, and necessitating the development of new techniques in computer programming to perform the aperture synthesis mapping.

Today, not much has changed, with modern radio interferometers requiring large computing clusters to correlate, calibrate, and image the necessary data products. In Chapters 2 and 3, we have seen that the foundations of radio interferometery and aperture synthesis mapping can be reformulated to allow new methods of imaging with commensurate computational benefits.

However these new approaches do not represent a silver bullet that will cure the computational woes of radio interferometry. Indeed looking back at over seven decades of aperture synthesis techniques, we can see many techniques have been tried, but we continue to be up against the limits of modern hardware.

Progress in the field is made with taking the best examples of hardware and software we have today and building algorithms suited for them. These new algorithms are often predicated by advances in understanding the mathematical foundations of radio interferometry.

The arrays that are being designed and are slowly coming online in the present day are allowing much greater sensitivity and angular resolution than before. The major example at present is the Square Kilometer Array (SKA) (Dewdney et al. 2009a). This array will allow sensitivity at the level of μJy's, and with baselines on the order of 3000 km for SKA-Mid, provide exceptional angular resolution. This presents itself as a significant imaging problem, as it allows images on the order of 50000^2 pixels, which takes up well over 300GB of memory, if the images are represented in double precision floating point numbers.

The SKA-Low array, which is a low frequency array composed of phased array beamformed tiles of antennas, has similarly large baselines, but with a lower frequency. The main issue with SKA-Low in particular, although this issue is shared by SKA-Mid to a smaller extent, is the enormous quantity of visibility data products produced, which must be accurately calibrated and imaged. Whilst they are not held to the real-time requirements that we discussed in Chapters 2 and 3, the planned computing architectures of the SKA do not release visibility products to end users (Bolton et al. 2019). This means that the SKA computing backbone will shoulder the burden of responsibility in producing aperture synthesis maps. Necessarily this means producing them quickly and with high accuracy as is demanded by an instrument of this ambition and scale. Unlike other arrays which merely produce the visibility data products, often with some fairly coarse processing to 'flag' out RFI, the SKA will produce visibilities, and then image, calibrate and deconvolve the resulting images. This procedure significantly increases the demands on the computing infrastructure.

Within this, there is a particular element which we will consider in this chapter which serves to have deleterious effects on an instrument of this sensitivity: wide-field imaging. With sensitivities on the order of $1\mu Jy$, errors on the order of 1% in a map are devastating and will limit the overall science output. Such errors are especially significant in surveys where a large portion of the celestial sphere is imaged in a single map. These errors come from the w-term in Equation 1.16, with a discussion on this issue already undertaken in Chapter 3.

In this Chapter, we continue to expand on this wide-field imaging problem but within the context of the interferometry measurement equation pertaining to visibilities. We first state the major three problems that serve to frustrate imaging of the latest planned radio arrays:

1) Very Large Image Sizes — The computational and memory requirements of image sizes with a leading dimensional size in 10,000's of pixels is extraordinary.
2) Near Real-Time Imaging — Whilst not held to strict real-time requirements as with EPIC, due to not necessarily releasing visibility data products to the end user, images must be formed quickly and accurately from each observation using limited compute resources.
3) Wide Field Effects — With planned arrays incorporating baselines thousands of kilometres in length, this necessitates incorporating wide-field effects on large baselines, subject to the curvature of the earth, and large fields of view.

None of these points present an insuperable challenge but, when taken in combination, they are overwhelming. Either an existing extremely capable algorithm must be extended to support imaging in these instruments, or in absence of that, new techniques must be devised.

Any such algorithm must be able to take advantage of contemporary high performance computing infrastructures and techniques. Imaging schemes exist which serve to cope with wide-field effects such as w-projection (Cornwell et al. 2008) and w-stacking (Offringa et al. 2014), as discussed in Chapter 3. These schemes go some way towards solving the degradation in image quality from the w term in an efficient manner. Recent work has revolved around incorporating hardware accelerators such as GPUs, for example in the scheme set out in Romein (2016).

An exciting recent demonstration of a wide-field imager is using Image Domain Gridding Tol et al. (2018). This is a scheme which uses a combination of innovations from fourier theory, and a computational implementation where GPUs and locality are taken advantage of to greatly increase the accuracy of wide-field maps compared to incumbent techniques. The lynchpin of this technique follows a similar philosophy to Chapter 3 where the underlying mathematical assumptions are interrogated and re-imagined in a way that is amenable to a high performance implementation, discussed in Veenboer et al. (2017), utilising GPUs.

The crux of this scheme is using the fourier multiplication convolution theorem to allow the "gridding" operation, where visibilities are placed onto a regular lattice so that an FFT can be performed, to be done in the image space using a fast discrete fourier transform kernel on a GPU, and then multiplication by a convolutional taper function. This imaging scheme has a higher floating point cost on paper, but as discussed and shown in Chapter 3, the commonplace wisdom of this leading to lower performance is flawed. However in the end, the throughput is similar to existing wide-field imaging techniques, with greatly increased accuracy.

In this chapter, we present a new method of wide-field imaging: w-towers, which parallelises the imaging into a set of towers on a grid; this can be thought of as a variant of the w-stacking algorithm. We present a high performance implementation, and compare performance to incumbent techniques.

Also, we present work building on great advances in the state of affairs of imaging from the usage of optimised convolutional kernels, whose mathematical formalism was devised by Sze Tan in his PhD Thesis (Tan 1986). This uses a mathematical least-squares optimisation of the convolutional gridding function, nominally a PSWF, to gain much greater image accuracy than any existing technique. This has recently been repopularised in Ye (2019), with the demonstration of a modified form of *w*-stacking, Improved *w*-stacking, which incorporates the optimal convolutional functions.

We go on to illustrate an improved iteration of *w*-towers, learning from the drawbacks discovered in *w*-towers. This Improved *w*-towers uses the Improved *w*-stacking algorithm for calculating each "tower" in this scheme. The Improved *w*-towers algorithm uses a novel distribution and recombination scheme to parallelise the wide-field imaging equation, overcoming the difficulties found in the original *w*-towers algorithm. This allows the scheme to be parallelised across a distributed computing cluster, which has become a much researched area in recent times, such as in Pratley et al. (2019).

This work was done in collaboration with Peter Wortmann, a member of the Cavendish Astrophysics group at Cambridge. Peter's contribution is the original *w*-towers parallelisation scheme, and its successor recombination scheme capable of distributing the wide-field imaging across a modern high performance computing cluster.

My own contribution (James Kent) is building the original high performance implementation of *w*-towers on the GPU, which led to an interest in resolving the work distribution problem. Within the framework of Improved *w*-towers I worked on the degridding problem to create a highly optimised CPU and GPU implementation for the problem of degridding, as well as formal validation of the redistribution theories.

This Chapter is divided into a tale of two gridders: *w*-towers in Section 6.2, and an efficient implementation of Improved *w*-towers which is described in Section 6.3. Within *w*-towers, the theory of the original *w*-towers implementation is discussed in Section 6.2.1, with the implementation and results described in Sections 6.2.2 and 6.2.3, respectively. A short discussion of outcomes is presented, which resulted in the switch to develop the Improved *w*-towers imager.

The theory of Improved *w*-towers and the reason for developing this from the original *w*-towers is discussed in two parts. In Section 6.3.1 the theory of optimal convolutional gridding functions is discussed with respect to a better method of *w*-stacking: Improved *w*-stacking. Then in Section 6.3.2, the redistribution framework is discussed which overcomes the shortcomings of the original *w*-towers, with proofs of the major mathematical results facilitating this technique. A highly optimised implementation of degridding using the optimal convolution kernels on a CUDA GPU is presented in Section 6.3.3. This highly optimised degridder slots

into the Improved w-towers framework as the main computational step. Finally results and performance for a simulated dataset are discussed in Section 6.3.4.

6.2 *w*-Towers

6.2.1 Theory

We can split Equation 1.17 into a convolution between the visibility in the u, v domain and a function, referred to as the w-kernel across the u,v, and w domains. Tim Cornwell formalises this in his w-projection papers Cornwell et al. (2008). Let $V(u, v, w)$ be a visibility on a continuous 3D domain as defined in Equations 1.17 and 1.16.

$$G(l, m, w) = \exp\left[2\pi i w(1 - \sqrt{1 - l^2 - m^2})\right] \tag{6.1}$$

$$\widetilde{G}(u, v, w) = \int_{-\infty}^{\infty} \exp\left[2\pi i\left(1 - \sqrt{1 - l^2 - m^2}\right)\right] dl\, dm \tag{6.2}$$

$$V(u, v, w) = \widetilde{G}(u, v, w) * V(u, v) \tag{6.3}$$

where $*$ represents a convolution operation. The sky brightness distribution can be generated by an inverse fourier transform of the visibilities convolved with this G function:

$$I(l, m, n) = \int_{-\infty}^{\infty} \left[V * \widetilde{G}\right](u, v, w) \exp\left[2\pi i ul + vm + wn\right] du\, dv\, dw \tag{6.4}$$

where $n = 1 - \sqrt{1 - l^2 - m^2}$. Here we can use the convolution theorem to implement the convolution with $G(u, v, w)$ after the fourier transform as a multiplication, instead of computing a costly and difficult to parallelise convolution, in the same way as w-stacking (Offringa et al. 2014). Thus the w-function is combined with the visibilities in an image space multiplication:

$$I(l, m) = \sum_{w_i = w_{min}}^{w_{max}} \left[\int_{-\infty}^{\infty} V(u, v, w_i) \exp\left[2\pi i\left(ul + vm\right)\right] du\, dv \right] \cdot G(l, m, w_i) \tag{6.5}$$

where w_i represents a particular w-value for a w-plane spaced by some Δw. Thus $V(u, v, w_i)$ represents the set of visibilities that apply to a particular w-plane. We thereby shift the burden of w-correction onto the density of the w-planes that we use, such that more closely placed w-planes, defined in Equation 6.5, result in a more accurate final dirty map. Too many w-planes in our w-stack will be costly to compute; this is offset by each visibility being convolved with a small w-kernel, $\widetilde{G}(u, v, w)$ to correct for the mismatch between a visibility and its closest w-layer:

$$I(l, m) = \sum_{w_i = w_{min}}^{w_{max}} \left[\int_{-\infty}^{\infty} \left[V(u, v, w_i) * \widetilde{G}(u, v, \delta w)\right] \exp\left[2\pi i\left(ul + vm\right)\right] du\, dv \right] \cdot G(l, m, w_i) \tag{6.6}$$

where δw represents the mismatch in w between the visibility and its closest w-plane, w_i.

The adaptation in w-towers is to split the problem in u and v as well as in w, thus defining multiple w-stacks, hence w-towers. In w-stacking the large grid size leads to slow FFT's due to cache misses because of poor data locality in memory. In w-towers, the grid is split into chunks as shown in Figure 6.1.

Each tower is a set of w-planes, and that chunk is calculated by doing the inverse fourier transform to image space, adding to the data already calculated, then multiplying by the w-kernel to get to the next w-plane. Mathematically each chunk should be of the form:

$$I\prime(l\prime,m\prime) = \sum_{w_i=w_{min}}^{w_{max}} \left[\int_{\infty}^{\infty} \left[\hat{V}(u,v,w_i) * \widetilde{G}(u,v,\delta w)\right] \exp\left[2\pi i\big(ul + vm\big)\right] du\, dv \right] \cdot \widetilde{G}(l,m,w_i) \tag{6.7}$$

where $\hat{V}(u,v,w_i)$ corresponds to all visibilities that exist in a volume bounded by the u and v coordinates of the w-tower base, along with being bounded by the extent covered by the w-plane in the w co-ordinate axis. The total number of w-planes, N, is a function of the total w range, and the increment between each layer:

$$N = \frac{w_{max} - w_{min}}{\Delta w} \tag{6.8}$$

where Δw is a user defined parameter. Each of these chunks is added to the final grid in its correct position, then a single inverse fourier transform on the large grid is executed, which takes us to our final image by Equation 1.27.

There is a small overlap between each tower base that corresponds to the support size of the w-kernel in the u and v domain. This is an improvement over existing algorithms in terms of computational efficiency. By confining activity to small chunks (with some management for the overlap between chunks) this presents a fairly trivial parallelisation problem. It will be shown that w-towers can be parallelised fairly simply without costly synchronisation mechanisms, using OpenMP and CUDA.

The w-towers algorithm is very similar to w-stacking, but instead of the subgrids representing each w-layer being the size of the whole grid, w-towers partitions the grid horizontally, and only keeps a single layer per partition in memory at one time. This is much more efficient in terms of memory usage, locality, and exposed parallelism. It rotates the parallelism axis, so instead of parallelising horizontally (through w) we parallelise in u and v, and sequentially process in w for each u, v partition. It is important to note that we do not necessarily need to have a serial tower per u, v partition. Bins can overlay each other provided they are not duplicating work on visibilities. In practice the w-towers scheme presented here provides a straightforward way of parallelising the measurement equation.

The use of smaller fourier transforms on each layer takes advantage of locality, as each chunk should fit into the L3 cache of the CPU, meaning the FFT can be very efficient. For GPU programming, the benefit is concurrency that facilitates all w-planes being gridded without the significant memory usage of w-stacking.

It is important to note that this breaks the strict fourier transform relationship in 1.26 and becomes an approximation to it. Motivated by the practical benefits explained above, w-towers and its derivatives are empirically motivated as they break from the strict adherence to the fourier relationship between the sky and the visibility measurements.

6.2.1.1 Complexity Analysis for w-towers

A complexity analysis is presented as a way of directly comparing algorithms, independent of implementation. This analysis allows the algorithms performance and scaling to be quantified as a function of independent parameters. These parameters are enumerated in Table 6.1.

The complexity has been analysed for w-towers and compared against w-projection, and w-stacking, in Table 6.2. Care has been taken to compute the approximate number of floating point instructions to allow a finer comparison, with the *caveat emptor* previously discussed of using floating point operations as predictor of performance. The subscripts translate as follows:

Table 6.1: Parameters for complexity analysis.

Symbol	Meaning
N_V	Number of Visibilities
N_G	Grid Size
N_{SG}	Sub-Grid Size
N_S	Support(Kernel) Size
N_C	Number of Chunks
N_F	Number of Floors

where the support size is derived as such:

$$N_S = \left\lceil \left[\frac{w\theta}{2} \right]^2 + \left[\frac{\frac{w^{\frac{3}{2}}\theta}{2}}{\pi n} \right] \right\rceil^{\frac{1}{2}} \tag{6.9}$$

where w represents the w-increment to correct for, θ is the field of view size for the observation, and n represents where the w-kernel is truncated. For all subsequent analysis, this truncation value is set at 0.01. This formula is from Mitchell and Bernardi (2014).

From here, the chunk size can then be derived, where the margin size between chunks is set to the kernel size:

$$N_C = \frac{N_G}{N_{SG} - N_S} \tag{6.10}$$

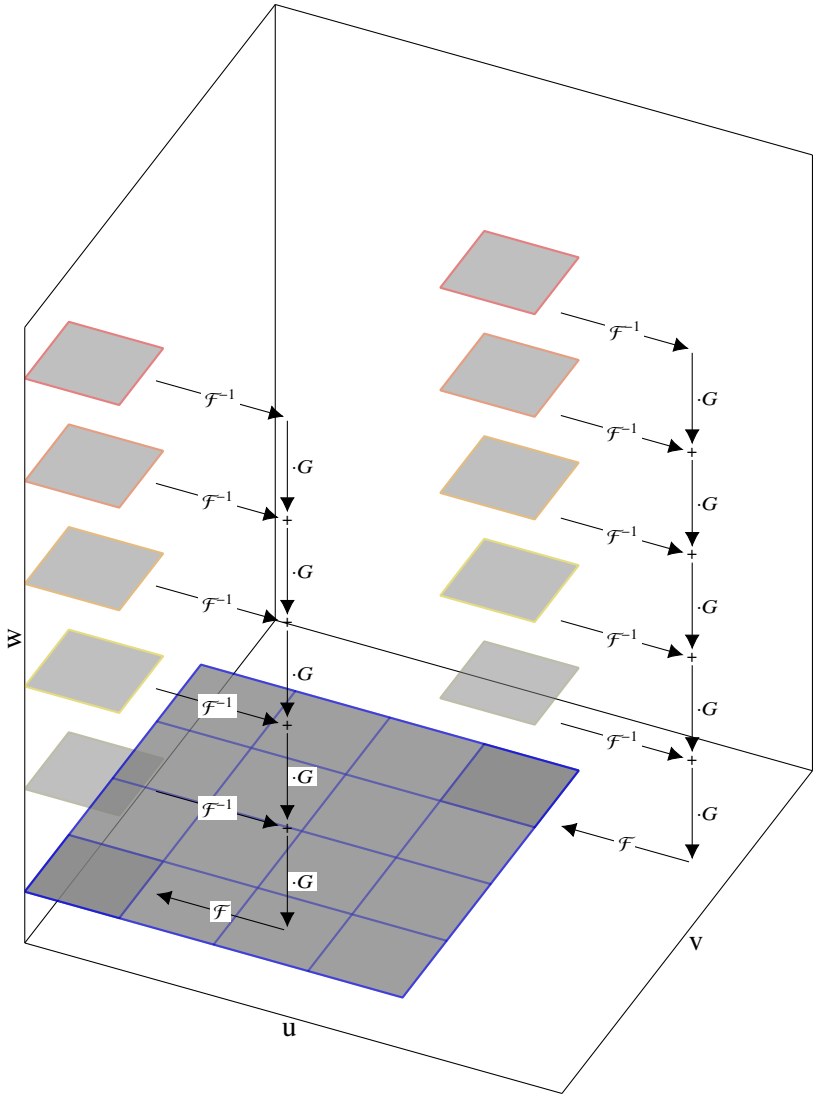

Figure 6.1: Visual Representation of w-towers parallelisation scheme. Visibilities are gridded on each w-plane within a w-tower, and an FFT is performed back to the image space and then the w-transfer to the next w-plane is performed and an FFT is performed to visibility space where additional visibilities are gridded, and this is done iteratively until all w-planes are covered. Each visibility is convolved with a small w-projection kernel in practice to decrease the number of w-planes required. After all towers are finished processing, they are added back to the main grid at $w = 0$ and a final FFT is performed to image space, giving the dirty map.

The power of *w*-towers comes from shifting the complexity into elements that are more amenable to efficient computation, such as smaller Fourier transforms and Hadamard products which cache more easily in memory, due to smaller grid sizes compared to *w*-stacking and *w*-projection.

Table 6.2: Complexity analysis for *w*-towers versus *w*-stacking and *w*-projection.

Operation	*w*-towers	*w*-stacking	*w*-projection
Gridding	$6N_V N_S^2$	$6N_V N_S^2$	$6N_V N_S^2$
Fourier Transforms	$5N_G^2 log N_G^2 + N_C N_F N_{SG}^2 log N_{SG}^2$	$5N_F N_G^2 log N_G^2$	$5N_G^2 log N_G^2$
w-correction	$6N_C N_F N_{SG}^2$	$6N_F N_G^2$	N/A
w-transfer	$18N_C N_{SG}^2$	N/A	N/A
Reduction	$2N_C N_{SG}^2$	$2N_F N_G^2$	N/A

In terms of complexity, *w*-towers and *w*-stacking are almost identical, with the gridding, fourier transforms, and *w*-correction being of equivalent complexity. The *w*-transfer is different, because in *w*-towers the tower is processed from top (the *w*-plane at largest positive w value), to bottom, and the final subgrid needs to be transferred back to the $w = 0$ plane for correct imaging.

This does mean however that the final reduction operation is of comparatively lower complexity in *w*-towers, as compared to *w*-stacking. *w*-towers is a faster imaging scheme for a large number of visibilities, with the addition of being able to compensate for large FFT's as image sizes increase by creating better locality through small FFT's.

The implications of the total FLOP count, and how the algorithms scale for different parameters, paints a complicated, but navigable picture. We set the following parameters for a tutorial analysis:

Parameter	Value
Grid Size	2048
Visibilities	10^4
Support Size	16
Margin Size	8

In Figure 6.2 we can see the scaling as a function of subgrid size. It is clear that *w*-stacking is more efficient unless the *w*-towers parameters are not chosen carefully. At low sub-grid sizes and at specific larger subgrid sizes, the FLOP count of *w*-towers is lower than for *w*-stacking. The reason for this is shown in Figure 6.3 where the flop count for the subgrid FFT's are plotted alongside a graph showing the ratio of the subgrid size to the number of chunks required to cover the grid size stated previously. Subgrid size needs to be chosen to minimise spurious FLOP's. Unfortunately at the larger subgrid sizes, the sizes which allow a FLOP count below that of *w*-stacking sit at subgrid sizes which are not a power of two, leading to inefficient FFT's,

where FFT's benefit most (speed-wise) from power of two input space sizes (Frigo and Johnson 2005).

Another important parameter to look at is the scaling as a function of the number of floors partitioning our Euclidean space, in the w-axis. In Figure 6.4 we can see that w-towers improves on w-stacking in this respect, meaning it should scale more efficiently for wide field imaging and high declination phase-centre imaging, where the range in w can be very high.

Figure 6.5 shows how the two approaches vary for visibility input size, with w-towers enjoying a marginal overall advantage. This is because of the cost saving in w-correction and Fourier Transforms, however in a large visibility limit they are all drowned out by the gridding cost. The effects of techniques such as baseline dependent averaging (Wijnholds et al. 2016), could be important here, however our cost is based on the raw number of visibilities inputted into the imager.

The memory usage is also profiled. The memory used is very implementation and optimisation specific, so calculating theoretically is only sensible to a point but there are trends that can be identified. The memory cost for the major contributors to overall memory cost are shown in Table 6.3.

Table 6.3: Memory cost as function of time.

Operation	w-towers	w-stacking	w-projection
Visibilities	$40N_V$	$40N_V$	$40N_V$
Grid Size	$16N_G^2$	$16N_G^2N_F$	$16N_{SG}^2N_C^2$
w-Correction	$16N_{SG}^2$	$16N_G^2$	N/A

It is clear that w-stacking has a memory problem compared to other gridding algorithms, as it is designed and implemented. As the w-term becomes more prominent in a dataset, and the number of floors increases, the memory cost of w-stacking quickly spirals, with the assumption that all w-planes are kept in memory at one time. Indeed, w-towers could be susceptible to this if the 'towers' were split up to facilitate additional parallelism. In its current implementation, the memory cost stays relatively bound, and has few drawbacks compared to other gridders. There is a commensurate increase in memory locality due to parallelising in this way, which helps increase performance.

6.2.2 Implementation

In this Section, we describe an implementation of the original w-towers parallelisation scheme, which led eventually to the idea of Improved w-towers. The w-stacking scheme facilitates low overhead parallelism by distributing computing as a function of w, with w-towers aiming to make further improvements by parallelising in u and v.

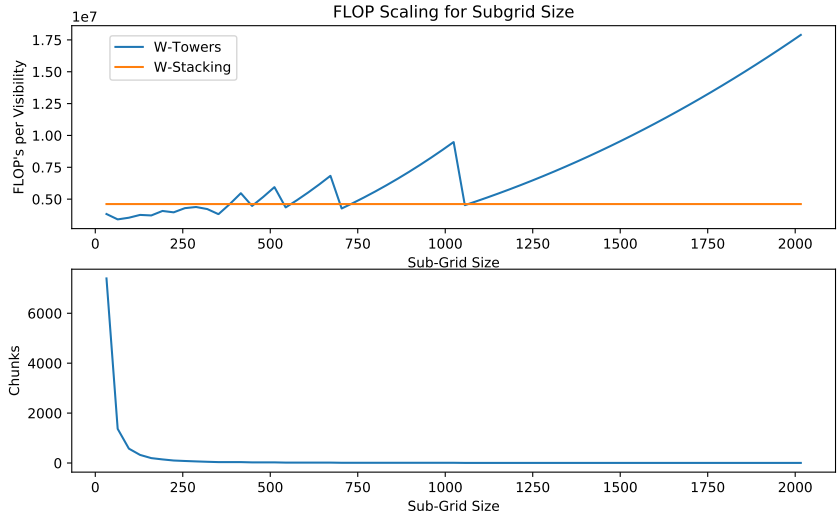

Figure 6.2: *w*-towers scaling as a function of subgrid Size in *w*-towers. *w*-stacking plotted for comparison.

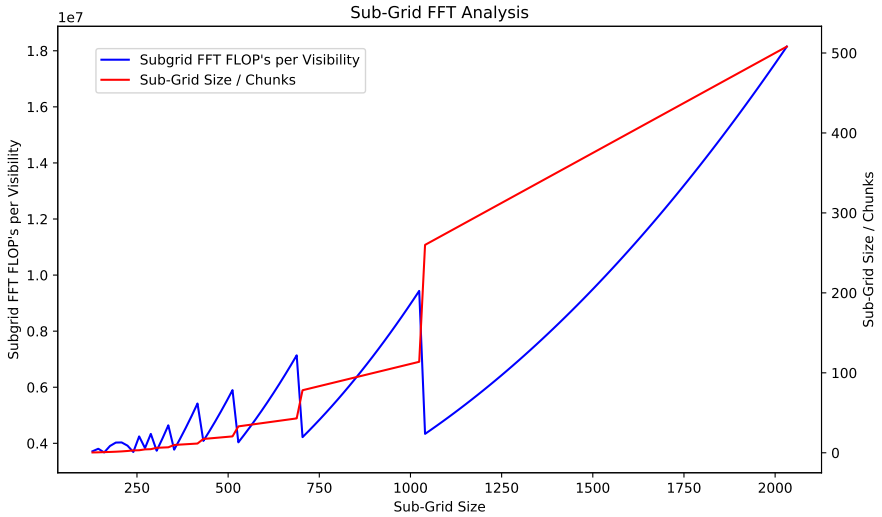

Figure 6.3: *w*-towers scaling as a function of subgrid FFT's. Ratio of subgrid size and chunk count is plotted to explain the FFT curve. Main grid size is 2048^2

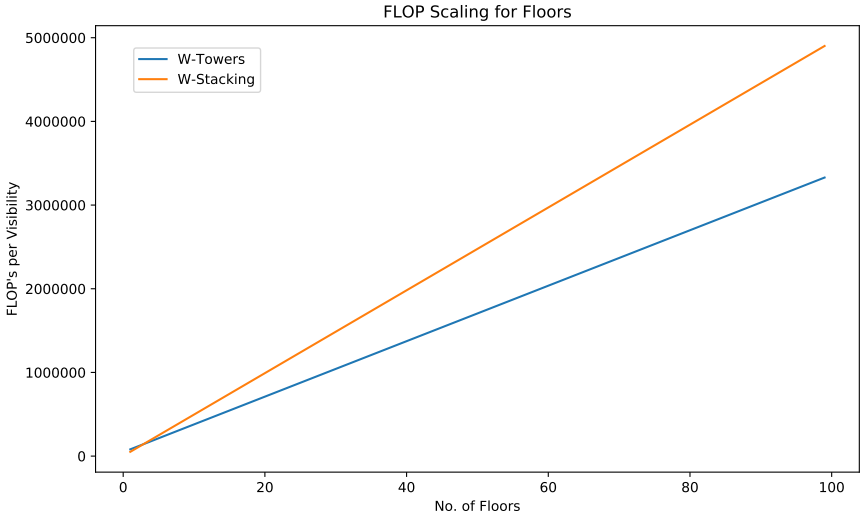

Figure 6.4: *w*-towers scaling as a function of number of floors. Each floor is distributed at a set interval in *w*. *w*-towers sub-grid Size is 128.

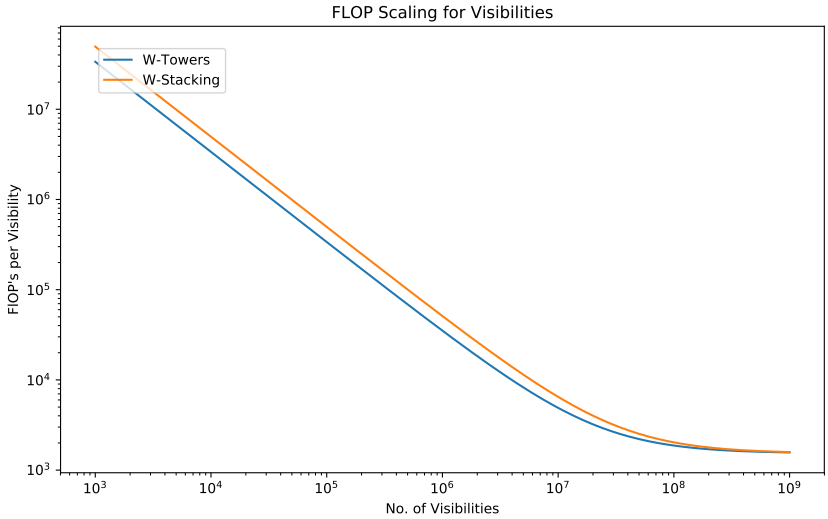

Figure 6.5: *w*-towers scaling as a function of visibility number. Number of floors is 100. Sub-grid Size is 128.

The implementation avoids explicit thread synchronisation mechanisms as much as possible. The only synchronisation mechanism used involves atomic instructions within the convolutional Romein gridder (Romein 2012).

6.2.2.1 OpenMP

The source code for *w*-towers and the parallelised version is not written in this thesis because of its length, however it is available on request from the author, and is submitted alongside this thesis.

The crux of the parallelisation is declaring an array of pointers to different subgrids which are all allocated at the same time. Each subgrid represents one chunk of the *w*-towers grid. The loop arithmetic is then defined to be one-dimensional, instead of two-dimensional which is what you would expect for a function operating on a 2D domain. Then an OpenMP pragma statement is used of the form:

```
#pragma omp parallel for schedule(dynamic) private(cy,cx)
for (chunk = 0; chunk < (grid_size / chunk_size + 1) * (
   grid_size / chunk_size + 1); chunk++) {
...
```

This statement indicates to the OpenMP engine that the FOR loop is to be parallelised with a dynamic scheduler, which is one of the scheduling modes of OpenMP. This means that as a thread finishes its current chunk it is assigned a new chunk to work from. This allows the workload to be balanced across all of the different threads. This parallelism without synchronisation means the threads are independent of each other up until the final addition to the main grid.

6.2.2.2 CUDA

A reference GPU implementation is provided using the NVIDIA CUDA framework. The parallelism is oriented in the same way as OpenMP; we parallelise by chunk.

The execution model of a CUDA GPU is now described briefly. The GPU is split into several streaming multiprocessors, where each has compute resources available, split between CUDA cores. These support 32-bit and 64-bit floating point computation, as well as integer computation, in addition to shared on-chip memory, which has a significantly higher bandwidth than the global graphics card memory. This memory can be thought of as analogous to the cache on a CPU. The multiprocessors are independent of each other, and there is no communication between different segments of the algorithm except for shared memory.

Each workload is represented by kernels, which are split into blocks; these are further split into a user defined number of threads, with each thread then operating on its assigned elements. Whereas a CPU uses up to several dozen threads on a single node, a single GPU can have tens of thousands of threads "in flight". The grid assigns blocks for the GPU to work on; these are the coarse elements of work, and the blocks map threads to CUDA cores. Threads are assigned in groups of 32, known as a Warp. Multiple blocks can execute concurrently if resources are available for it on each multiprocessor; this would be handled by the driver for the GPU. This is known as the Single Instruction Multiple Threads (SIMT) paradigm.

The *w*-towers scheme lends itself nicely to GPU parallelism using the asynchronous execution framework provided in the CUDA API; each tower can be represented by a CUDA stream, and all of the workloads for each chunk are sent to the GPU at the same time. The driver then assigns work to the GPU from all of the streams concurrently. The driver is optimised to keep the GPU saturated with work, overlaying kernels from different streams where possible.

A pseudocode of the main work loop of the program is below:

```
for(stream = 0; stream < total_chunks; ++stream){

 for(w_plane = 0; w_plane < wp_max; ++wp){
   Grid <<< dimGrid, dimBlock, 0, Streams[stream] >>> (param);
   cuFFT();
   fresnel <<< dimGrid, dimBlock, 0, Streams[stream] >>> (param);
 }
transfer_w0 <<< dimGrid, dimBlock, 0, Streams[stream] >>> (param);
final_fft();
}
```

More prior work is required compared to *w*-projection, but the advantage is in the simplicity of the towering step, and the smaller kernels. The main prior step necessary is binning the data in 3D to ensure that each scatter gridder only receives visibilities that correspond to its bounds. This means we don't need to make time consuming comparisons in device code.

This implementation goes to serve an initial proof of concept implementation of *w*-towers incorporating the optimised convolution scheme in Romein (2012). As will be discussed, there is a limitation in the *w*-towers parallelisation scheme which prompted to a revisiting of the original underlying principles. This led, together with recent advances in optimal convolution theory, to a major iterative advance in the procedure.

6.2.3 Performance

Tests were ran for speed and accuracy. Two datasets were used:

- An SKA-Mid Dataset simulated with OSKAR (Mort et al. 2010). Phase centre at the zenith. The w values range between -400 and +400.
- A VLA 'A' configuration dataset, at a high declination phase centre, resulting in a w value range between -3500 and +3500. The dataset copy is a 7x7 point source square.

The results were collected on Ubuntu 16.04 LTS, with dual six-core Intel Xeon E5-2630 processors, and 64GB of DDR3 random access memory (RAM). For the CUDA implementation, preliminary tests were done on the development machine with an Nvidia Kepler K20C. Further tests were performed on Nvidia Pascal P100's on Wilkes2, part of the Cambridge Centre for Data Driven Discovery (CSD3), comprising the GPU counterpart to Peta4.

6.2.3.1 OpenMP

When using an SKA-Scale dataset with a test card dataset imaged at the zenith, with significant visibility averaging, w-projection takes 33-35 seconds on our test system, and using about 10 Gigabytes of RAM in the process. With w-towers, this can be reduced to just over 4 seconds, with only moderately increased memory usage over a single threaded implementation. It is important to balance the chunk and margin size with system specifications like available cache size. Failure to do this can be result in significantly longer running times.

In Figure 6.6 the execution time is measured as a function of the number of threads used in the main gridding stage. The scaling is very good with a typical asymptotic roll-off. Further gains can be realised with parallelising some of the binning and pre-gridding stages of the algorithm. Whilst the VLA Dataset is on a significantly smaller grid it does have a far higher number of visibility points, thus providing more work for the gridder.

The SKA Dataset has the benefit of being on a very large grid, thus meaning there is a very large number of sparse subgrids to work on. However it doesn't have particularly high w-values, with a range of 400. As w-towers is intended for wide-field imaging with a nominally high w-range, a test dataset using the VLA 'A' Configuration was imaged. This is shown in Figure 6.7.

An exploration of the parameter space with smaller chunk sizes and margins is shown in Figure 6.8 Smaller chunk and margin sizes speed up the gridding process, with the optimum being a margin space which is big enough to account for the w-Kernel convolution. Anything more is surplus.

Based on the VLA execution results, a Chunk Size of 200 and a margin of 40 was chosen. The results for the execution time versus w-increment are shown in Figure 6.9 for the chosen

minimum in our previous results. The picture when looking at operational intensity becomes less clear. The chunk and margin size is plotted against operational intensity in Figure 6.10. Operational Intensity is calculated as so:

$$Operational\ Intensity = \frac{FLOPs}{Cache\ Line\ Size \times LLC\ Misses} \tag{6.11}$$

where $LLCMisses$ refers to cache misses in the lowest level cache (L3 in an Intel GPU), necessitating a fetch from main system memory. The cache line size in an Intel CPU is 64 bytes long. The FLOPs and LLC Misses are provided by the performance counters present on most mainstream CPU architectures.

Seemingly as execution time increases operational intensity increases, although intuition would dictate that we expect operational intensity to *decrease* as execution time increases. It would appear that performance counters are not reliable in this respect. The greater speedup as a result of decreasing subgrid size, whilst coming slightly from a reduced complexity, cannot be explained by that effect alone. It must therefore be the benefit of the locality, where small subgrids exist primarily in the CPU cache, compared to main memory.

The comparison with algorithms such as w-projection is significant. The implementation of w-towers can complete a grid of the VLA data in 80 seconds whereas w-projection can take 200 seconds using variable w-kernels, significantly more than w-towers. For a fair comparison, the w-increment for w-towers is set to the step-size of the w-projection kernels, 20 in this case. A single thread is used for these two cases to assure fair comparison. Memory usage is 3.53 GB for our w-projection gridder, but only 2.81 GB for the w-towers grid. This is because of the larger w-Kernels required for w-projection.

w-towers memory usage will increase slightly as the ratio between chunk size and margin size decreases. For example, with a chunk size and a margin size of 80, the max allocated memory is 3.00GB, compared to 2.81GB for a margin size of 20. In light of the significant performance costs to running with high margin costs, without any clearly shown gains, this is not a problem.

6.2.3.2 CUDA

The performance compared to our OpenMP implementation is shown in Table 6.4. We use Gigapoint Additions per Second (GAPS) to illustrate how fast points are being added to the grid. Additional tests were performed on an Nvidia P100 GPU, shown in Table 6.5. With the P100, the entire dataset was imaged due to greater available memory. The convolution kernel for the w-projection gridder was set at 16x16, with 8x oversampling.

On the P100 cards the w-projection gridder becomes significantly more efficient, which greatly boosts performance compared to Kepler. This is likely because of the addition of a

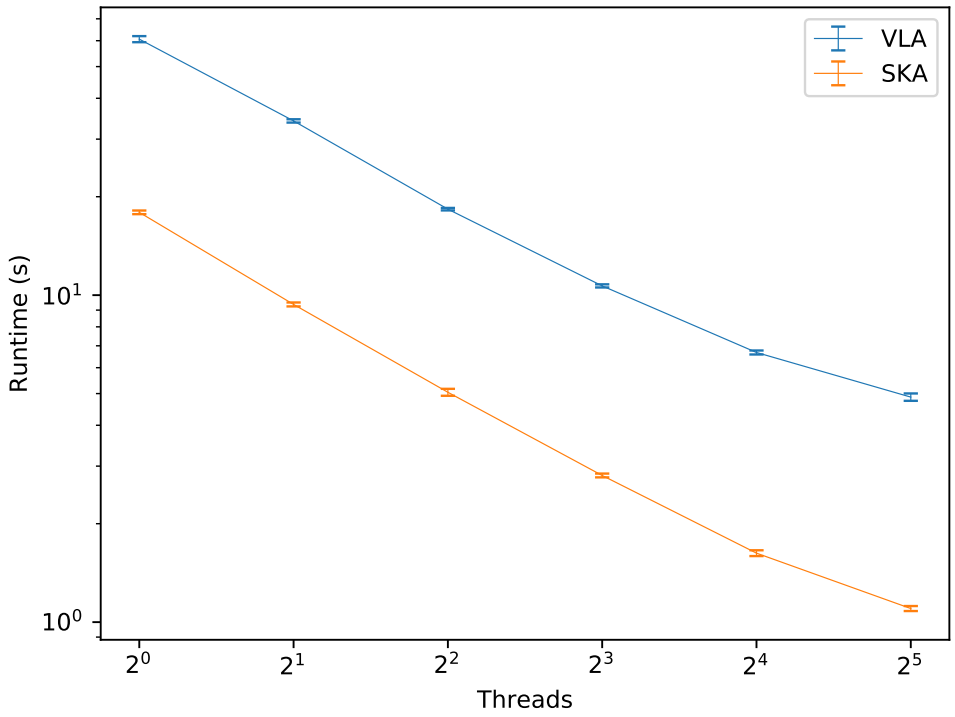

Figure 6.6: Execution Time vs Number of Threads for the SKA and VLA datasets. This is with a *w*-kernel of 20 pixels along the leading dimension, with an oversampling of 8. The chunk size is set at 100. The SKA dataset has undergone visibility averaging as a pre-processing step, thus ends up with less visibilities (130816 visibility records) than the VLA dataset (702000 visibility records).

Table 6.4: Measured Execution Times for *w*-towers on a Kepler K20C GPU versus CPU. Grid size 1024^2. Kernel size of 16x16.

		w-towers	*w*-projection			*w*-towers	*w*-projection
Time(s)	CUDA	5.43	3.65	GPAS	CUDA	1.23	1.87
	CPU	3.34	N/A		CPU	1.98	N/A

Table 6.5: Measured Execution Times for Full VLA Dataset on Pascal P100 GPU.

	w-towers(16x16)	*w*-projection(16x16)	*w*-projection(150x150)
Time(s)	5.77	2.54	238.513
GPAS	3.5	7.086	6.764
Memory Used (GB)	6.35	4.51	5.74

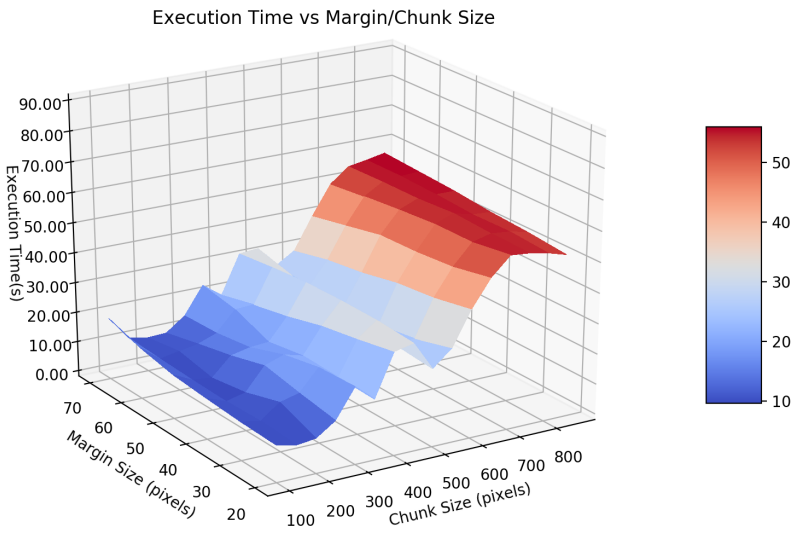

Figure 6.7: A measure of execution time vs chunk/margin size for a simulated VLA Dataset. The w-increment between w-planes is 20.

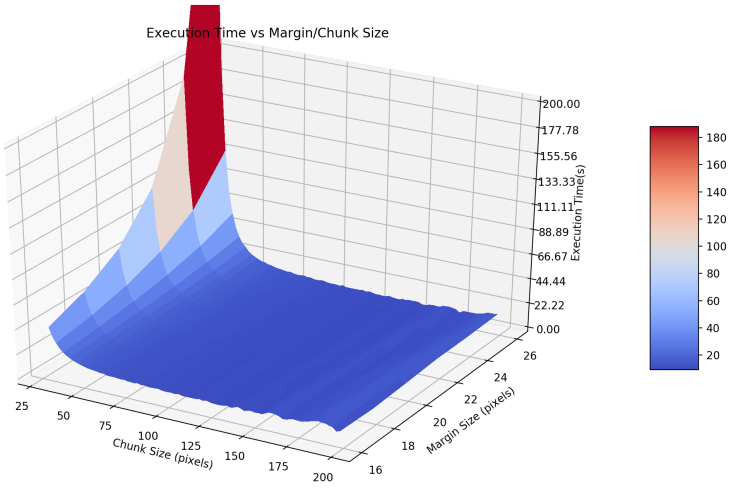

Figure 6.8: Execution time for VLA Dataset for smaller chunk/margin sizes.

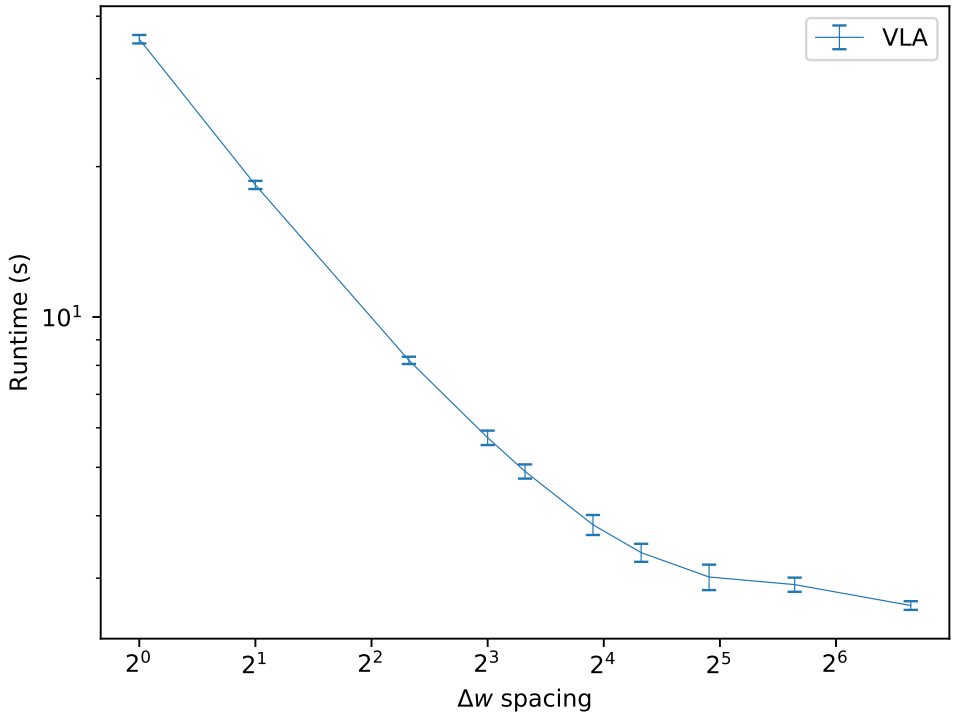

Figure 6.9: A measure of execution time vs w-increment in the w-towers algorithm, with the VLA Dataset. The VLA dataset was used because of its very high w-dependence. This is benchmarked over 10 iterations of each Δw value, with 32 threads. Each w-tower chunk size is 100 pixels in u and v. The w-kernel is sized at 20 pixels in 1D, with an oversampling of 8.

hardware atomic double precision floating point addition, which is not available on previous generations of CUDA enabled GPUs. The memory usage varies, but w-towers uses more memory due to the binning required to distribute visibilities across the 3D u, v, w space. The w-projection method has not been optimised to use variable kernels in this instance. This will be easiest done when the gridder is extended to allow for aw-projection (Bhatnagar et al. 2008), which needs to be done on a per baseline basis. From tests in OpenMP using variable w-Kernels, the expected w-projection time should see a reduction by 60%.

The requirement for binning in 3-dimensions, combined with multiple scatter grid kernels, which take up 90%+ of the time according to the CUDA profiler, indicates this implementation of w-towers is not viable due to the prohibitive cost of gridding. This matches with what we expect from the complexity analysis; at some point, gridding cost overpowers everything to the level that w-projection is by far the fastest way of imaging. The complexity benefits of smaller FFT's as shown in our previous analysis, do not carry over to GPUs due to the design of cuFFT,

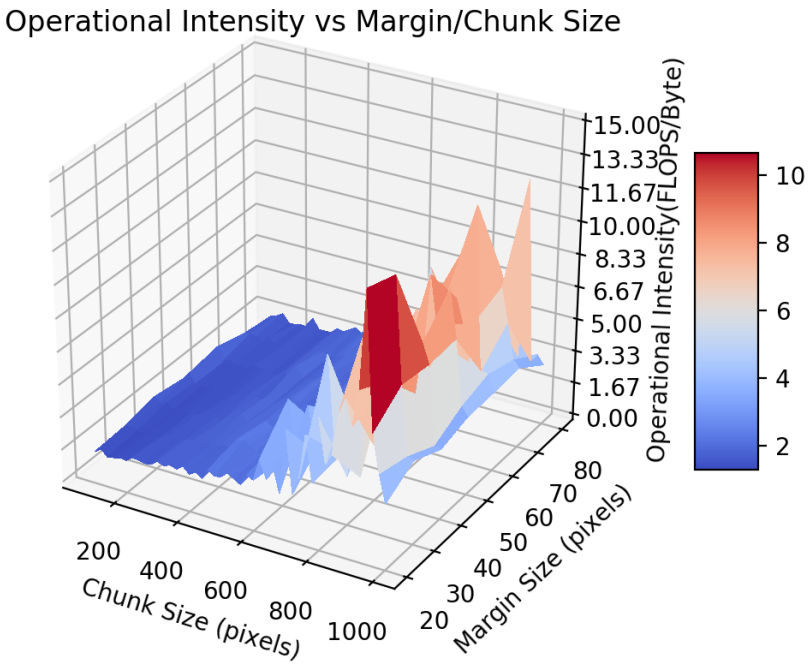

Figure 6.10: A measure of operational intensity vs chunk/margin size for a simulated VLA DAtaset. Winc is 20.

and is therefore not viable in this context in this implementation. To get more performance out of CUDA and GPUs for imaging, complete removal of w-projection is the way forward and this is explored in Section 6.3.

Furthermore, execution time is inversely proportional to subgrid size, which is the opposite of what we see on the OpenMP version. Thus locality benefits seen on a CPU do not carry over. Indeed cuFFT is very efficient for large matrices on CUDA (Kent and Nikolic 2016). This is likely due to larger sizes of fourier transforms being able to saturate the memory bus between the main graphics card memory and the GPU more efficiently.

6.2.3.3 Accuracy

To compare accuracy, the mean square error(MSE) was used:

$$MSE = \frac{\sum_{l=0}^{N}\sum_{m=0}^{N}(I\prime[l,m] - I[l,m])^2}{N^2} \tag{6.12}$$

where $I\prime[l,m]$ is the DFT calculated directly. A fast DFT algorithm, written in CUDA, is used for this purpose, because of the prohibitive cost of calculating the DFT for the large visibility

datasets used here. The source code for the DFT is included alongside the *w*-towers code. By using this method, residuals were created and the MSE calculated, and the results shown in Table 6.6. It can be shown that *w*-towers is equivalent, if not slightly better for smaller *w*-kernels than *w*-projection.

Table 6.6: Mean Squared Error for *w*-towers and *w*-projection.

Gridder	Mean Squared Error
w-towers	9.43 $\times 10^{-}5$
w-projection	9.74 $\times 10^{-}5$

The configuration of *w*-towers has an impact on accuracy. For example, Figure 6.11 shows the residuals between two *w*-towers images. The centre of the grid has very low error, with error increasing towards higher l, m as is common in wide-field imaging. The algorithm is numerically stable, with MSE between images created with the same parameter being of the order of 1e-24.

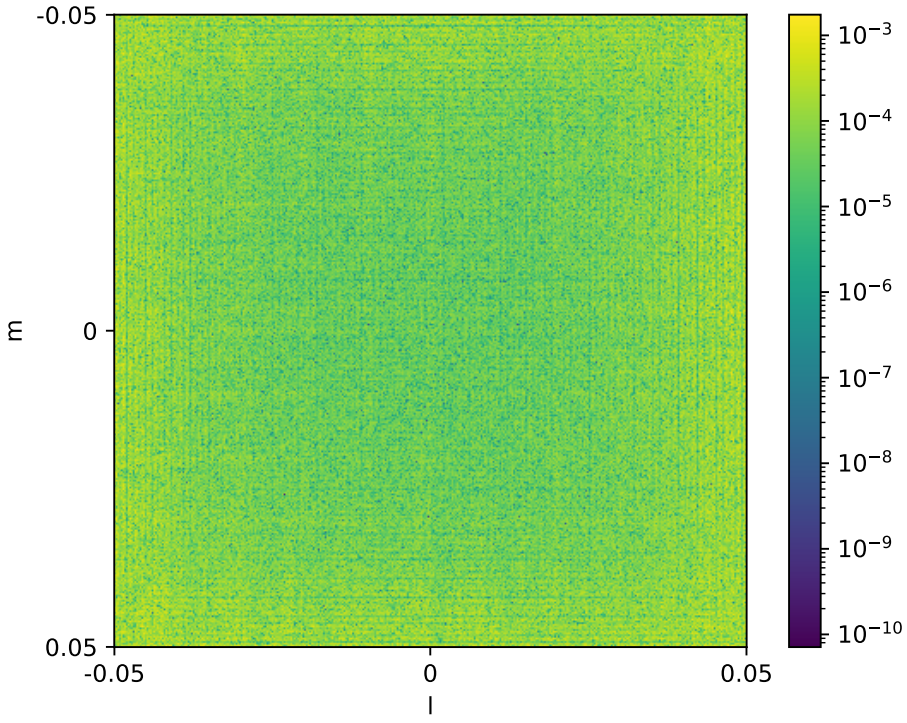

Figure 6.11: Residuals plotted between two *w*-towers images, Chunk Sizes 100 and 400 respectively. Margin is 20, and Winc is 20. The units are in arbitrary values.

Residuals between *w*-towers are plotted at *w*-Increments of 1 and 5 respectively in Figure

6.12. We notice that there is more wide field error at the edges which is expected due a higher w-increment having less power in correcting these wide field errors.

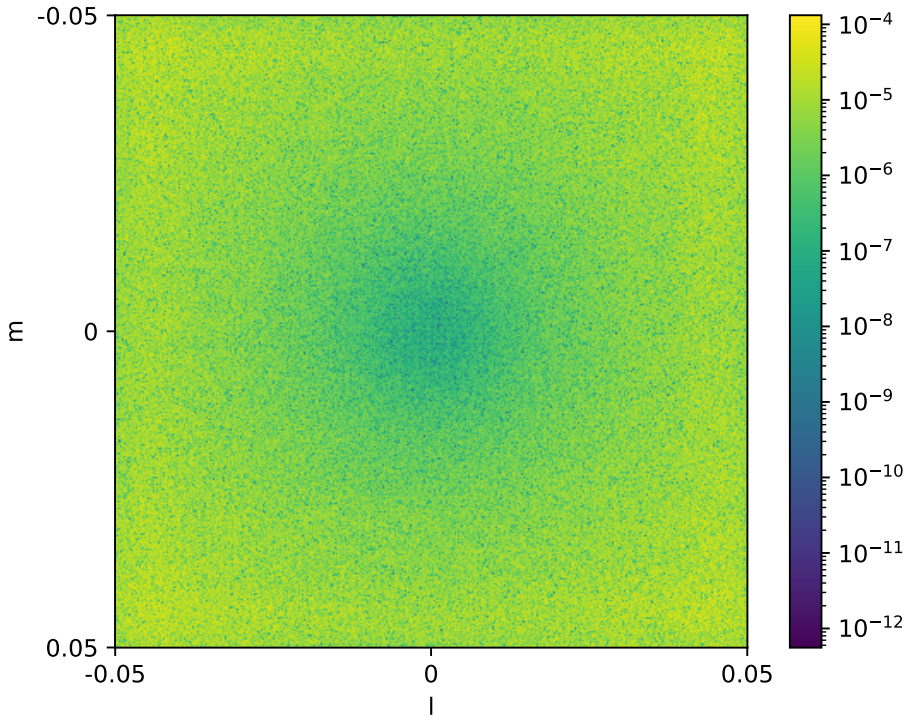

Figure 6.12: Residuals plotted between two w-towers Images. w-Increments of 1 and 5 respectively. The units are in arbitrary values.

6.3 Improved w-Towers

6.3.1 Optimal Convolution Theory and Improved w-stacking

We move on from describing the w-towers algorithm to describing an evolution of this approach which concentrates on parallelising the imager in such a way that computation is not restricted to a single node.

The interferometry measurement equation is calculated using the van-Cittert Zernike theorem, as discussed in Chapter 1. One of the necessary issues that arise from this is that Equations 1.26 and 1.27 are costly to compute directly for a visibility dataset of any reasonable size.

This computation can be made more computationally tractable by interpreting Equations 1.26 and 1.27 using a Fast Fourier Transform (FFT) algorithm, which greatly improves the

scaling of computing the Fourier Transform from $O(n^2)$ scaling to $O(nlogn)$. This requires putting values onto the regular grid defined by the FFT. The requisite gridding was discussed in Chapter 2 and Section 6.2, and apart from being computationally tricky to implement itself, it adds mathematical complications that need to be corrected for.

The first complication is that the FFT algorithm defines a transform between regular fourier pairs of $\mathbb{C}^N$ dimensionality spaces. These paired vector spaces are discretely sampled as defined by the Discrete Fourier Transform, with the FFT an efficient method of calculating the DFT for regularly sampled functions such as Equations 1.26 and 1.27. In radio interferometry this means that visibilities (or electric fields such as in Chapter's 2 and 3) are gridded onto a two dimensional flat regularly spaced grid. Baseline vectors are defined in terms of (u, v, w) with the w component being negligible only for a completely flat array with an observation phase centre at zenith. Thus the flat grid assumption for the FFT fails in many real-world observation scenarios.

Equations 1.16 and 1.17 can also be interpreted in terms of a 3D fourier transform, where the only valid sky pixels exist where the celestial sphere intersects with each voxel in the 3D regular grid. This is very costly to implement in practice, however, and will not be discussed further.

Another complication that results from using a regularly spaced grid in $\mathbb{C}^N$ is that each visibility measurement is defined on a continuous domain, whereas the FFT is a purely discrete domain; visibilities will almost always never lie directly on a grid point. This mismatch between a visibilities true position and the position of the nearest grid point is solved with a convolution between the visibility and an anti-aliasing function, most often a prolate spheroidal wave function (PSWF). However there still exists an error, or misfit, between the image calculated from the FFT and the image calculated from the DFT. This is a consequence of an non-optimal anti-aliasing function, henceforth known as the Grid Convolution Function (GCF), and also the lack of perfect w-correction.

A brilliant innovation occurred in Tan (1986), whereby both of these issues were corrected using a least-squares minimization between the dirty map as defined by the Discrete Fourier Transform, and the model convolutional kernel. This has recently been extended, with orientation to practical implementation, in Haoyang Ye's doctoral thesis Ye (2019), and a recent article (Ye et al. 2019). The interested reader is advised to start with Tan (1986) for the fundamentals, and move on to Ye (2019) for modern developments with this technique including the improved w-stacking technique which will be discussed here.

The crux of this scheme is an examination of the convolutional gridding scheme, firstly in one dimension, as a least-squares misfit between the dirty map resulting from the application

of the FFT, and that of the DFT:

$$\Delta(l) = \left[I(l)_{FFT} - I(l)_{DFT}\right]^2 \tag{6.13}$$

where:

$$I(l)_{DFT} = \sum_{i=0}^{N} \omega_i V_i(u) \exp\left[2\pi i u_i l\right] \tag{6.14}$$

$$I(l)_{FFT} = h(l) \sum_{i=0}^{N} \sum_{j=-C/2}^{C/2} c(j-u_i)\omega_i V_i(u) \exp\left[2\pi i u_i l\right] \tag{6.15}$$

Here, ω is a weighting term commonly used in radio astronomy to downweight the contributions from short baselines to even out the relative strength across different sky modes corresponding to broad emission versus narrow emission. The variable V represents individual visibility samples, and $h(l)$ and $c(u)$ represent the gridding correction function in image space and the gridding convolution function in visibility space. The variable C is the integer support size of the GCF. These symbols are chosen to harmonise with the notation in Tan (1986) and Ye et al. (2019).

Equation 6.13 can be re-arranged with the usage of the Cauchy-Schwarz inequality, $|a.b|^2 \leqslant |a.a| \cdot |b.b|$ to give:

$$\Delta(l)^2 \leq \sum_{i=0}^{N} \left[\omega_i V_i(u)\right]^2 \sum_{i=0}^{N} \omega_i \left|1 - h(l) \sum_{j=-C/2}^{C/2} c(j-u_i) \exp\left[2\pi i (j-u_i) l\right]\right|^2 \tag{6.16}$$

where we now define:

$$L(l,u) = 1 - h(l) \sum_{j=-C/2}^{C/2} c(j-u) \exp\left[2\pi i (j-u) l\right] \tag{6.17}$$

Equation 6.17 can be re-written in the same way as Equations 2.18 and 2.19 in Tan (1986) by defining the fractional offset of u as $\eta = u - \lfloor u \rfloor \in [0,1]$:

$$L(l,\eta) = 1 - h(l) \sum_{j=-C/2}^{C/2} c(j-\eta) \exp\left[2\pi i (j-\eta) l\right] \tag{6.18}$$

$$\hat{L}(l) = \int_0^1 \left|L(l,\eta)\right|^2 d\eta \tag{6.19}$$

Equation 6.19 is what is defined in the literature as the "Map Error Function", and forms the analytical function to be minimised for accurate reconstruction of the sky brightness distribution using the FFT. The total error to minimise can be defined as per Equation 2.22 of Tan (1986):

$$E = \frac{1}{2x_0} \int_{-x_0}^{x_0} \hat{L}(l) dl = \frac{1}{2x_0} \int_{-x_0}^{x_0} \int_0^1 L(l,u)\, du\, dl \tag{6.20}$$

where x_o represents a fraction of the overall sky map to optimise over. For the examples in this Chapter, x_0 is nominally 0.5. The minimisation is performed by differentiating Equation 6.20 with respect to h and c and then minimising E over both h and c jointly; this is the method used in Tan (1986). Alternatively, as set out in Ye (2019), we can solve the set of simultaneous equations formed by differentiating with respect to c, to determine the inverse Toeplitz matrix of the solutions. This is the method that is used here to generate our gridding functions, using the Python package scipy. After deriving $c(u)$ through this method, $h(l)$ can be derived by Equation 2.21 of Tan (1986):

$$h(l) = \frac{\hat{c}(l)}{\sum_{j=-\infty}^{\infty} \hat{c}(l-k)^2} \tag{6.21}$$

where $\hat{c}(l)$ can be derived as the real component of the fourier transform of the grid convolution function $c(u)$:

$$\hat{c}(l) = \int c(u)\cos(2\pi u l)du \tag{6.22}$$

Since the gridding problem, and fourier transforms in general, are defined on an orthonormal set of basis functions, it is quite simple to construct higher dimensional grid convolution functions, such as in $\mathbb{R}^3$ capturing the u, v, w degrees of freedom of a real radio interferometer. This is because the convolution functions are in practice seperable for u, v and w. A rigorous proof of the seperability of the convolution functions is shown in Proofs 1 and 2 of Chapter 2 of Tan (1986). This has obvious advantages in that several 1D functions can be tabulated instead of the more standard method of computing a non-seperable 2D w-projection kernel which requires tabling in memory due to the inseperability of the kernel.

Using the minimisation scheme above, the Grid Convolution Functions and Grid Correction Function can be derived numerically. This is done for a u/v GCF of support size 8, and for a w GCF of support size 4. The kernels are oversampled at a rate of 4096 points per grid point. These u/v and w Grid Convolution Functions are shown in Figures 6.13 and 6.14. The corresponding Grid Correction Functions are shown in Figures 6.15 and 6.16. A prolate spheroidal wave function is plotted for comparison.

6.3.2 Improved w-Towers Theory

The parallelised scheme discussed in Section 6.2, was a major advance, but it proved difficult to speed up in practice due to a logjam in computational load close to the origin in the visibility domain, as a result of a large number of visibilities from short baselines. This causes the chunks that cover the central portion of the u/v grid to shoulder the major burden of the computational load. There is the additional difficulty that this scheme does not facilitate parallelisation across several computer nodes, due to the requirement for a final large FFT.

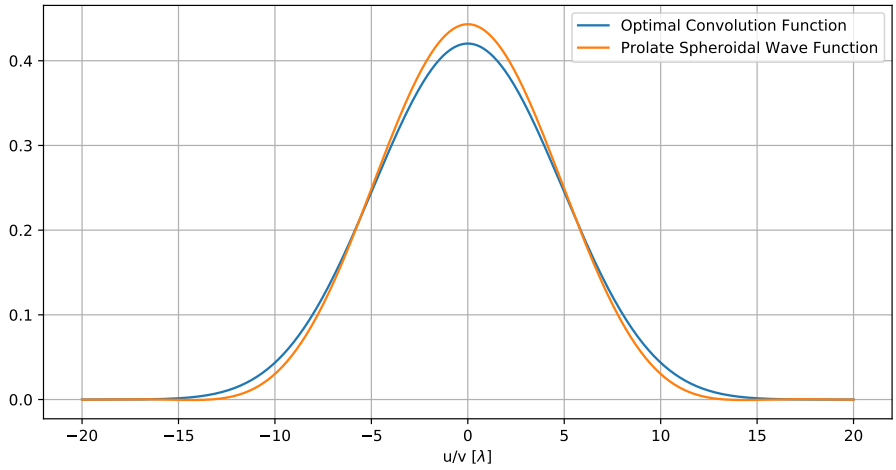

Figure 6.13: An optimised grid convolution function for the u, v dimensions for a support size of 8, oversampled at a rate of 4096. A prolate spheroidal wave function is plotted for comparison. The Field of View is set to a l/m size of 0.1. x_0 is 0.5.

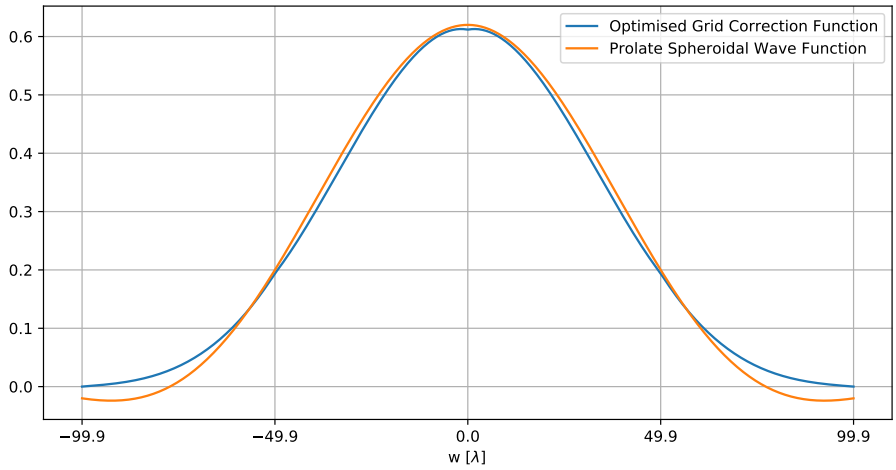

Figure 6.14: An optimised grid convolution function for the w dimension for a support size of 4, oversampled at a rate of 4096. The co-ordinate basis is based on the optimal spacing of the w-planes. The optimal spacing here is Δw=49.93. A prolate spheroidal wave function is plotted for comparison.

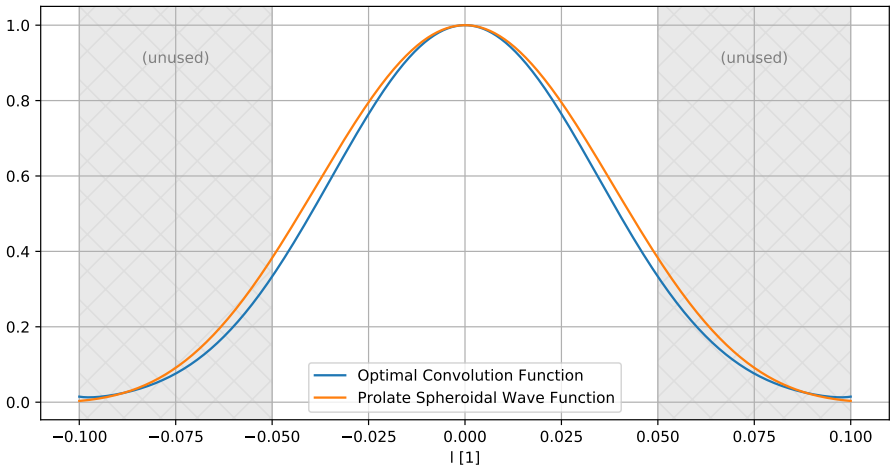

Figure 6.15: An optimised grid correction function for the l, m dimensions for a support size of 8, oversampled at a rate of 4096. A prolate spheroidal wave function is plotted for comparison.

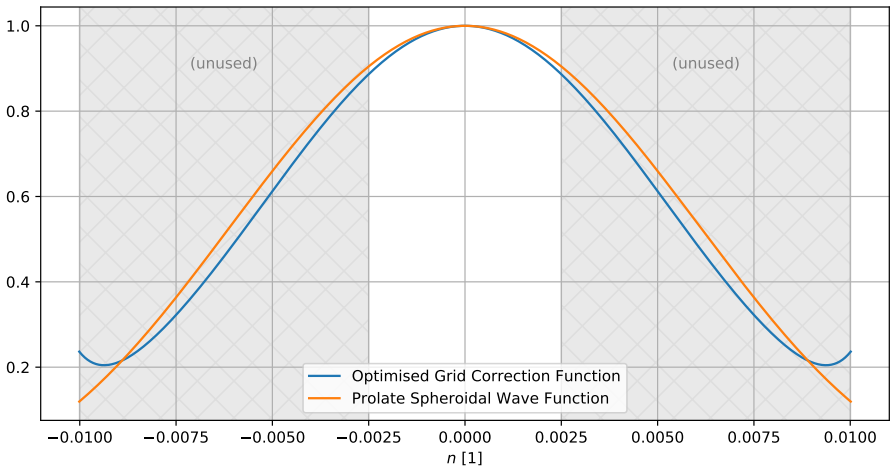

Figure 6.16: An optimised grid correction function for the n dimension for a support size of 4, oversampled at a rate of 4096. A prolate spheroidal wave function is plotted for comparison.

An alternative approach is to distribute the subgrids (or facets in the image domain) across a cluster, so that each node can have a dynamic amount of workload assigned to it. This approach significantly complicates reconstruction of the resultant sky brightness distribution. A new scheme, proven here, is presented that allows the parallelisation of the interferometric measurement equation in a very loose and flexible manner. This scheme can be equally well interpreted as a sparse distributed approximated FFT. The optimised CUDA implementation presented in Section 6.3.3 sits within this overall framework.

This approach is possible by taking advantage of the linearity of the fourier transform, along with the multiplication-convolution theorem. This innovation is separate wholly of the physical relation described by Equations 1.16 and 1.17. Firstly, we consider the problem in 1D by defining a pair of dualled fourier spaces $I(l)$, analogous to the sky brightness distribution in one direction, and $V(u)$, the spatial coherence measurement:

$$V(u) = \int_{-\infty}^{\infty} I(l) \exp\left[-2\pi i u l \right] dl \tag{6.23}$$

$$I(l) = \int_{-\infty}^{\infty} V(u) \exp\left[2\pi i u l\right] du \tag{6.24}$$

Next, we consider all the subgrids, where a subgrid is a subset of the u domain. We likewise consider all the facets, where a facet is a subset of the l domain:

$$V(u)_i \subseteq V(u)\ s.t.\ V(u)_i = R(u)_i V(u) \tag{6.25}$$

$$I(l)_j \subseteq I(l)\ s.t.\ I(l)_j = I(l) \int_{-\infty}^{\infty} \hat{R}(u))_j \exp\left[2\pi i u l\right] du \tag{6.26}$$

Here, $R(u)$ and $\hat{R}(u)$ are defined in terms of rectangular functions and their fourier transforms:

$$R(u)_i = rect(u)_i = \begin{cases} 1 & \forall |u - u_i| \leq n_u \\ 0 & \text{otherwise} \end{cases} \tag{6.27}$$

$$\int_{-\infty}^{\infty} \hat{R}(u)_j \exp\left[2\pi i u l\right] dl = rect(l)_j = \begin{cases} 1 & \forall |l - l_j| \leq n_j \\ 0 & \text{otherwise} \end{cases} \tag{6.28}$$

where n_u and n_j define the individual subgrid and facet sizes, respectively. It now follows that:

$$V(u)_i = R(u)_i V(u) = R(u)_i \sum_j \hat{R}(u)_j * V(u) \tag{6.29}$$

$$I(l)_j = \int_{-\infty}^{\infty} \left[\hat{R}(u)_j * \sum_i R(u)_i V(u)\right] \exp\left[2\pi i u l\right] du \tag{6.30}$$

with the assumption of correct normalisation in the fourier transform between $I(l)$ and $V(u)$, satisfying Parseval's theorem. Therefore:

$$V(u) = \sum_i R(u)_i \sum_j \hat{R}(u)_j * V(u) \tag{6.31}$$

$$I(l) = \sum_j \int_{-\infty}^{\infty} \left[\hat{R}(u)_j * \sum_i R(u)V(u)\right] \exp\left[2\pi i u l\right] \, du \tag{6.32}$$

The original specification of the problem in Equations 6.25 and 6.26 can now be split up into $i_n \times j_n$ dependencies. However, the truncation of $R(u)$ and $\hat{R}(u)$ will lead to unacceptable error in the reconstruction specified by Equation 6.31. This error arises from the jump condition that now exists at each sub-image and facet edge. This leads to a an overall oscillatory element to the reconstruction that increases error. Mathematically, this is analogous to the transient solutions that exist in the mechanical problem of a forced oscillator parameterised by a Heaviside step function, soluble by means of a Green's function.

Theorem 1 (Facet Split)**.** $I(l)_j = \int_{-\infty}^{\infty} \left[\hat{R}(u)_j * \sum_i R(u)_i V(u)\right] \exp\left[2\pi i u l\right] \, du$

Proof of Theorem 1. First notice that:

$$I(l)_j = \mathbb{F}\big(\hat{R}(u)_j\big)^{-1} \cdot I(l) \tag{6.33}$$

where we use $\mathbb{F}$ to abbreviate the fourier transform:

$$\mathbb{F}\big(f(\square)\big)(\breve{\square}) = \int_{-\infty}^{\infty} f(\square) \exp\left[-2\pi i \square \breve{\square}\right] d\square \tag{6.34}$$

This can be re-arranged, given that $I(l) = \mathbb{F}\big(V(u)\big)^{-1}$, as:

$$I(l)_j = \mathbb{F}\big(\hat{R}(u)_j\big)^{-1} \cdot \mathbb{F}\big(V(u)\big)^{-1} \tag{6.35}$$

From linearity of the fourier transform we can therefore construct:

$$I(l)_j = \mathbb{F}\big(\hat{R}(u)_j\big)^{-1} \cdot \mathbb{F}\Big(\sum_i R(u)_i V(u)\Big)^{-1} \tag{6.36}$$

Finally, we use the multiplication-convolution theorem to give:

$$\mathbb{F}\big(f(a) \cdot g(b)\big) = \mathbb{F}\big(f(a)\big) * \mathbb{F}\big(g(b)\big) \tag{6.37}$$

$$\mathbb{F}\big(f(a) * g(b)\big) = \mathbb{F}\big(f(a)\big) \cdot \mathbb{F}\big(g(b)\big) \tag{6.38}$$

Thus:

$$I(l)_j = \mathbb{F}\big(\hat{R}(u)_j * \sum_i R(u)_i V(u)\big)^{-1} \tag{6.39}$$

Q.E.D.

Theorem 2 (Subgrid Split)**.** $V(u)_i = R(u)_i \sum_j \hat{R}(u)_j * V(u)$

Proof of Theorem 2. First, begin with:

$$V(u)_i = R(u)_i V(u) \tag{6.40}$$

By invoking linearity, as in the proof of theorem 1, we can define:

$$V(u)_i = R(u)_i \sum_j \mathbb{F}\left(\mathbb{F}\left(\hat{R}(u)_j\right)^{-1} \cdot I(l)\right) \tag{6.41}$$

Then, using the multiplication convolution theorem as in Theorem 1, it follows that:

$$V(u)_i = R(u)_i \sum_j \mathbb{F}\left(\mathbb{F}\left(\hat{R}(u)_j\right)^{-1}\right) * \mathbb{F}\left(I(l)\right) \tag{6.42}$$

$$V(u)_i = R(u)_i \sum_j \hat{R}(u)_j * V(u) \tag{6.43}$$

Q.E.D.

The key to reducing the error is by splitting $R(u)$ and $\hat{R}(u)$:

$$R(u) = r(u)r'(u) \tag{6.44}$$

$$\hat{R}(u) = \hat{r}(u) * \hat{r}'(u) \tag{6.45}$$

where $r'(u)$ and $\mathbb{F}\left(\hat{r}'(u)\right)(l)$ have greater support than $r(u)$ and $\mathbb{F}\left(\hat{r}(u)\right)(l)$, such that:

$$V(u)_i = r(u)r'(u)V(u) = r(u)V(u) = R(u)V(u) \tag{6.46}$$

We are now in a position to redefine Equations 6.29 and 6.31:

$$V(u)_i = r(u)r'(u) \sum_j \hat{r}(u)_j * \hat{r}'(u) * V(u) \tag{6.47}$$

$$I(l)_j = \mathbb{F}\left(\hat{r}(u)_j * \hat{r}'(u) * \sum_i r(u)_i r'(u)_i V(u)\right)^{-1} \tag{6.48}$$

These expressions can be approximated as so:

$$V(u)_i \approx r(u) \sum_j \hat{r}'(u)_j * r'(u)\hat{r}'(u)_j * V(u) \tag{6.49}$$

$$I(l)_j \approx \mathbb{F}\left(\hat{r}(u)_j * \sum_i r'(u)_i \left(\hat{r}'(u) * r(u)_i V(u)\right)\right)^{-1} \tag{6.50}$$

where we have taken $r'(u)$ and $\hat{r}'(u)$, the approximating masks inside the summations, this simplifies the final reconstruction, and empirically has not had a significant effect on the error of this approximate technique. The computational cost is slightly increased as there is an

additional multiplication or convolution step (they are interchangeable as discussed with an appropriate Fourier transform) but it is intended that the Σ's are parallelised across many nodes of a compute cluster, making this extra cost negligible due to working in sub-grids and facets.

No function exists that allows both $r(u)$,$\hat{r}(u)$ and $\mathbb{F}\big(r(u)\big)^{-1}$,$\mathbb{F}\big(\hat{r}(u)\big)^{-1}$ to have finite support in both the u and l domains. This is because of the nature of the fourier transform, which is a many-to-many linear mapping where a subset of points in one domain will contribute to all points in the paired domain. However a function can be chosen that *approximates* this requirement, and the incumbent function that facilitates this is the PSWF, discussed several times thus far. By increasing the support of $r'(u)$ and $\mathbb{F}\big(\hat{r}'(u)\big)$ the PSWF allows the large errors from the truncation of the individual subgrids and facets to be decoupled and reduced. This is very similar to the margins that were introduced with w-towers but with a more rigorous understanding of the error.

The greatest advantage of this scheme is that it completely decouples the individual subgrids and facets from each other, meaning that they can be processed in a distributed fashion across a high performance computing cluster, with a recombination operation defined by the $\sum$'s in the preceding derivations.

6.3.2.1 Extension to Higher Dimensions

In the context of Equations 1.17 and 1.16, u, v and l, m are formed from orthogonal basis vectors, defined with the standard inner product by $\langle u, l\rangle = 0$, $\langle l, m\rangle = 0$. By this means we can operate under the pretense of seperability due to the orthogonality of the fourier transform. Considering this directly:

$$V(u,v) = \iint_{-\infty}^{\infty} I(l,m) \exp\big[-2\pi i(ul+vm)\big]\, dl\, dm \tag{6.51}$$

$$I(l,m) = \iint_{-\infty}^{\infty} V(u,v) \exp\big[2\pi i(ul+vm)\big]\, du\, dv \tag{6.52}$$

The functions $I(l,m)$ and $V(u,v)$ are not seperable, but the facet masks $R(u)$ and $\hat{R}(u)$ (and their decomposed approximations in Equation 6.44) can be extended by virtue of seperability:

$$R(u) \rightarrow R(u,v) = R(u)R(v) \tag{6.53}$$

$$\hat{R}(u) \rightarrow \hat{R}(u,v) = R(u) \otimes R(v) \tag{6.54}$$

Mathematical induction might be used to generalise this scheme of approximation to arbitrarily high dimensions. The boundedness of the error incurred by the above approximations in higher dimensions is not known, however.

We have deliberately neglected the w problem within this scheme, as this can be capably handled by the Improved w-stacking approach with the optimal convolution kernels described

previously. The error in this scheme deployed in aperture synthesis imaging is dominated by two terms:

1) Recombination of the Facets and Sub-grids
2) The *w*-correction

The first of these is capably handled by the approximate recombination and reconstruction scheme designed here. The *w*-correction must be corrected to the same or better levels of accuracy to ensure it is not the limiting factor.

We are thus able to distribute a set of facets and subgrids across a compute cluster using this method, where each facet and subgrid is processed using the Improved *w*-stacking method for either gridding or degridding.

6.3.3 Implementation

Improved *w*-stacking was implemented within the context of the Improved *w*-towers framework. Here we only showcase a high performance implementation of the Improved *w*-stacking algorithm, a sketch of which exists in Ye et al. (2019), with the usage of the optimal convolution functions described previously. Improved *w*-towers requires a highly accurate grid and image (or predict for visibilities) step which is supplied by this implementation of Improved *w*-stacking.

This implementation is done using a combination of C++ code, with high performance GPU-based elements written in the CUDA framework from NVIDIA. Use of the Standard Template Library from C++ is incorporated so as to abstract elements and make handling the underlying data structures easier without decreasing runtime performance. Additionally type templating was incorporated to allow precision of the overall imager to be controlled by allowing user control of the precision of the floating point arithmetic on both the CPU and GPU.

Several major optimisations are made compared to the original *w*-towers implementation for both the CPU and CUDA code, and represents a major iterative change, with development of the overall scheme actively progressing.

The integration of the redistribution framework with Improved *w*-stacking is beyond the scope of this thesis and represents work to be done. Here we describe an optimised CUDA version of the Improved *w*-stacking originally described in Ye (2019). This original version was written in PYTHON and not suitable for imaging large datasets. All of the high performance kernels herein described are contained in Appendix C. The overall source code is available alongside this thesis. The optimal convolution kernels described in Section 6.3.1 are pre-generated and imported into the Improved *w*-stacking application.

6.3.3.1 CPU Implementation

A reference version of degridding has been implemented in C++ that is highly parallelised across all cores of a CPU and to maximise the memory bandwidth. The parallelisation is made using a parallel for loop construct using the OPENMP framework, which is a widely supported add-on to most C/C++ compilers.

The software simulates a random sky of point sources with user-configurable Field of View sizes as well as the number of points. We have used the methodology of optimising for the most general case, or rather, the worst case, where a random distribution of sky sources is degridded to a random distribution of visibilities. Of course in practice, each array may have some regularities that increase memory locality but this cannot be taken for granted. By optimising for the most general case we can increase performance for all configurations of observation. In reality it may be possible for telescope-specific optimisation of degridding "lines" of visibilities, as baselines track out lines in u, v space as a function of time and frequency.

The OPENMP parallelised code works on a per-thread basis; each thread (of which there are up to two per physical core on modern computing hardware) does a single degridding convolution. Each thread gets a set number of visibilities to process at a time. This is a configurable parameter that is fed to the OPENMP framework via a PRAGMA statement. A PRAGMA is a way by which source code can inform compiler decisions. Thus each thread gets roughly 1000 visibilities to degrid before degridding the next batch of visibilities.

To aid the parallelism, we also make usage of non-unit striding of the memory to nullify the nefarious effect known as "cache thrashing". In a modern CPU a cache is associative, with each line in a CPU's on-chip cache being tied to a number of physical locations in memory. Thus if an element is in a CPU cache in a particular cache line, but the program wishes to access and cache something in the main system memory, if that address in main system memory is associated with the same cache line, it will evict the current cache line and supplant it with the new one. This is particularly prevalent for arrays with a size that is a power of 2. This is unfortunately the case here because we wish to use power of two FFT sizes as they are the most efficient in practice (Kent and Nikolic 2016).

If the CPU needs to access both elements in memory, but they both map to the same cache line, there is a clash and elements are repeatedly swapped into and out of the cache line; this is known as "cache thrashing". It is inefficient in taking longer to get the required data element to process an instruction, and it also indirectly slows down other areas of the application by using memory bandwidth.

Here we use non-unit striding by defining a custom TEMPLATE CLASS for both 2D and 3D matrices, which allows the memory strides of each dimension to be specified. In addition to

this it also allows type templating so that a user can select arbitrary numerical precision for the input vectors. The TEMPLATE features of C++ make it extremely easy to generalise in this way. The FFT's are performed using the incumbent standard FFTW (Frigo and Johnson 2005), using the "guru" interface which allows the FFT's to be performed on an array with non-unit striding.

The convolution to implement the degridding is done in all three dimensions simultaenously; all planes in the *w*-stack being formed by rapid iterative multiplications with the fresnel pattern and fourier transforms. We take advantage of the seperability of the optimal convolution kernels as discussed above and in Tan (1986), to form the 3D kernel on the fly with no pre-loading of the entire convolution kernel required. After this the visibilities are degridded at their nominal locations.

6.3.3.2 CUDA GPU Implementation

The CUDA implementation takes advantage of the enormous parallelism that GPUs offer through the usage of the Single Instruction Multiple Thread (SIMT) paradigm. GPUs offer exceptional speed up for computationally bound algorithms which have a good surface area for parallelisation.

In the case of the improved *w*-stacking algorithm we are considering here (and often throughout radio astronomy), we are dealing with an algorithm that is memory bandwidth limited. GPUs nevertheless still offer significant speedup due to the type of RAM used. The RAM on GPUs (usually known as GDRAM) often has significantly higher bus widths, giving much greater memory bandwidth than is available on a CPU. Thus GPUs will still offer good speedup compared to a CPU with a suitable algorithm. The major downside however is that the execution model of a GPU is highly specialised and therefore requires specialised knowledge and experience to speed up these algorithms.

When this problem was first approached it was difficult to see how to best implement the Improved *w*-stacking algorithm onto the GPUs execution model in an efficient manner. The *w*-stacking functionality is verbatim the same as described above for the *w*-towers implementation, with some modification for the C++ templating scheme we developed for this method. The major target for optimisation is the degridding convolution kernel described already.

We originally sought to maintain the generality of the CPU implementation, whereby any size and combination of convolution kernels could be used, giving granular control of the accuracy. This was implemented in a similar scheme to the CPU implementation with each thread in a thread-block on the GPU being responsible for a single degridding convolution. This is then saved back to the main visibility array in GDRAM. A visual overview of this scheme is shown in Figure 6.17.

This scheme was found not to give particularly good performance in practice. It was decided to restrict the support size of the optimal convolution kernel to allow a CUDA kernel to be executed in the most optimal manner, maximising the utilised memory bandwidth of the CUDA GPU. To do this, a technique was used known as "warp parallelism". A CUDA GPU executes threads in batches of 32. These 32 threads are known as a warp; they are able to work concurrently within the warp, and also with other warps if necessary. We restricted the convolution size of the kernel to a support size of 8 in the u, v directions and a support size of 4 in the w direction. Since there are 32 threads in a warp, we can split the warp into four quarter-warps, each of 8 threads. Each quarter warp processes one of the 4 w-planes by iteratively adding along the v direction: the 8 threads of each quarter-warp process the 8 elements in the u axis first, and then the next 8 elements along the v axis are added, and so on until all the grid/convolution points are processed.

This suffices to gain the accuracy required for even the second iteration implementation of the SKA, with 1μJy sensitivity. Greater accuracy than this is possible with larger convolution kernels. Infact, the convolution kernels can, as shown in Ye (2019), offer mathematically perfect correction within the limits of rounding errors inherent to floating point arithmetic. Thus there is not much lost by restricting the convolution size: we maintain our ability to generate accurate sub-image facets within the Improved w-towers scheme whilst gaining the ability to aggressively optimise for performance.

After the warp parallelised accumulation of grid points is multiplied with the correct convolution value, an explicit warp synchronisation is performed and a warp reduction sum is performed. This sums the thread local registers over the CUDA warp. The final value is cached in shared memory until a final write can be performed to GDRAM. The caching in shared memory is performed so that eventually a coalesced write can be performed where the full width of the memory bus is exploited to optimise for memory bandwidth. A visual overview of this optimised warp parallelised version is shown in Figure 6.18.

6.3.4 Performance

In this section, we present the performance of the Improved w-stacking degridding algorithm. The optimised implementation of the Improved w-stacking scheme has been performed so that it can fit into the Improved w-towers framework that has been discussed, with the aim of alleviating both the accuracy and performance drawbacks of the original w-towers implementation.

Performance is specified in terms of the computational runtime of the full Improved w-stacking algorithm, as well as the accuracy of the method for different convolution kernel support sizes and oversampling. There is no difference between the accuracy of the CPU and CUDA implementations.

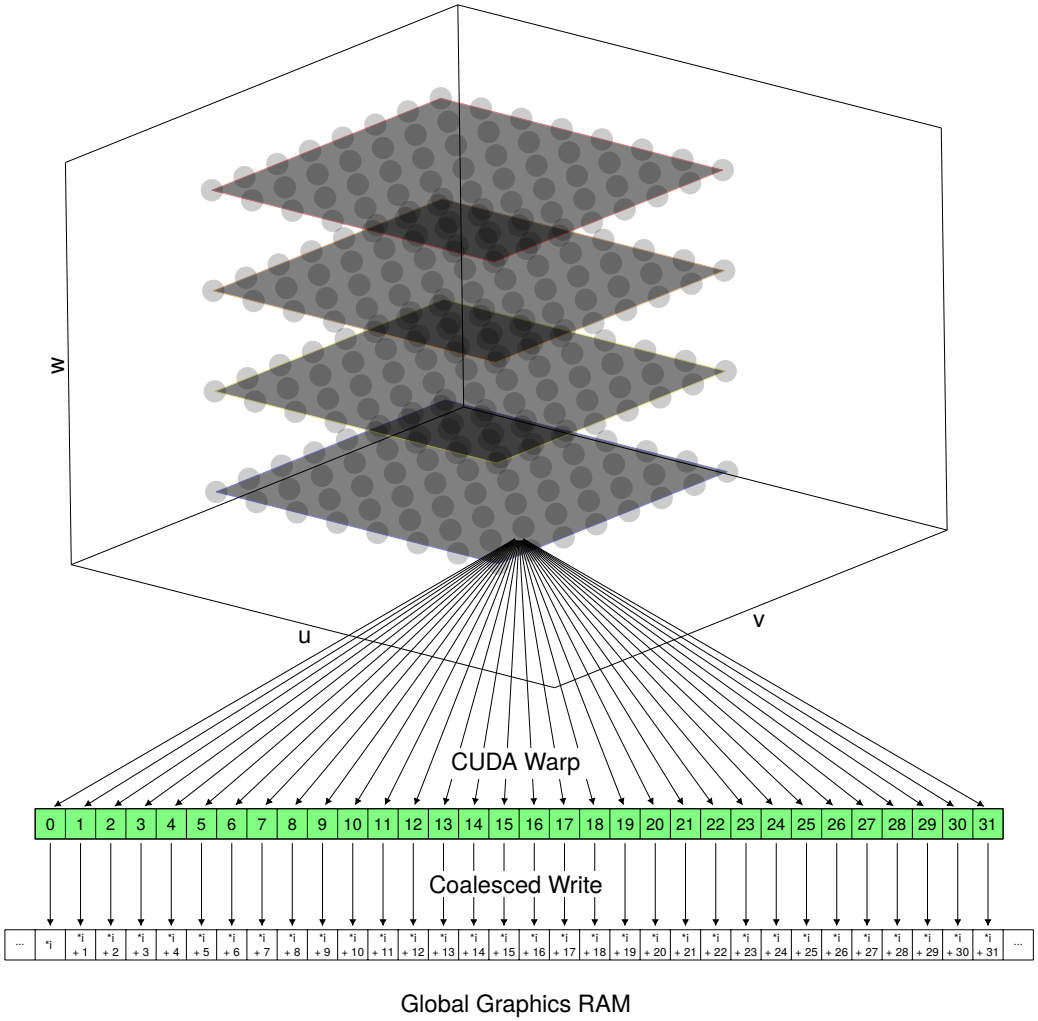

Figure 6.17: Visual Representation of original *w*-stacking degridding kernel for CUDA. This is based on directly mapping the parallelisation scheme used in the OpenMP C++ code to CUDA, which in practice leads to poor performance. A highly optimised version of this kernel is presented in Figure 6.18.

The performance results shown here are for a specific set of parameters, which are stated in Table 6.7. The optimised convolutional kernels were pre-generated for these parameters and were loaded into the program at start-up.

These tests were performed on the Centre for Data Driven Discovery (CSD3), constituting the main academic supercomputer of the University of Cambridge. Since the optimised code is intended to function as the leaf computation within the distributed Improved *w*-towers scheme, each test was performed on a single node. The CPU code was tested on a dual Intel Xeon Gold 6142 node of CSD3, with each CPU having 16 cores at a clock rate of 2.6 GHz, with a single thread per core. The CUDA implementation was tested on a single GPU of a single node of

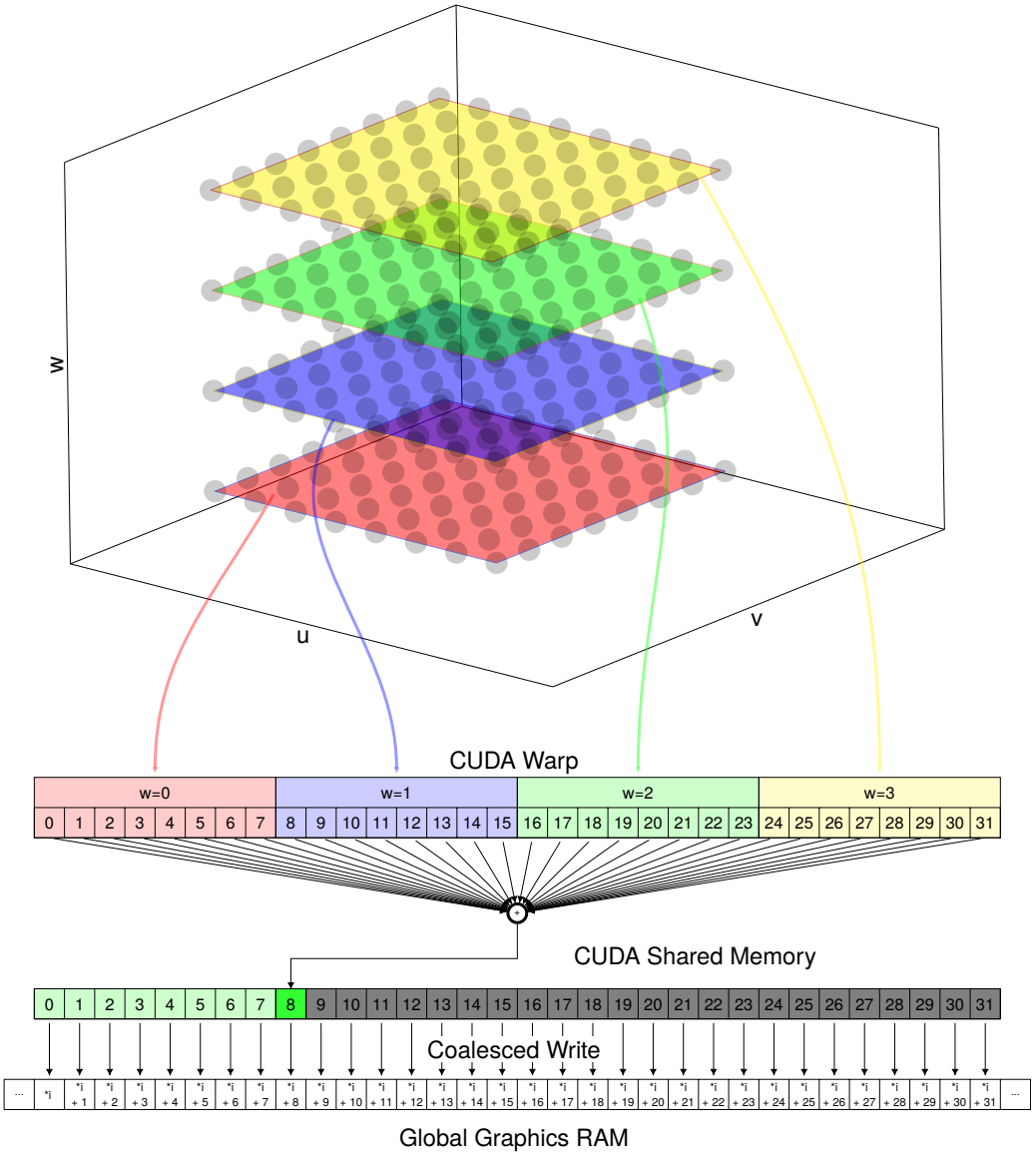

Figure 6.18: Visual Representation of Optimised Improved *w*-stacking CUDA Kernel. Each quarter warp is responsible for a *w*-plane within the 3D convolution. Each thread within a quarter warp multiplies a grid value with a convolution value along the coherent axes of memory and adds this value to a thread local register until all rows are processed, which allows coalesced reads from memory and optimised locality. A warp reduction is done after that, and the value stored in shared memory until a final coalesced write is made back to the main graphics card memory.

Parameter	Value
No. of Visibilities	15×10^6
Field of View	0.1 radians
Wavelengths Coverage in u/v	20480
Convolution u/v support	8
Convolution w support	4

Table 6.7: Parameters used in benchmarking the Improved w-stacking high performance implementation.

Wilkes-2, the GPU cluster component of CSD3. Each Wilkes-2 node is equipped with 4 Nvidia Pascal P100 GPUs.

6.3.4.1 Execution Time Performance

Table 6.8 shows the run-times for the execution of Improved w-stacking degridding. As expected, the optimised CUDA code is vastly superior to the performance on a CPU. This is a great improvement compared to the original w-towers computation. In particular the w-stacking code, which builds the discretised 3D (u, v, w) space, is extremely fast and efficient. This is due to the "embarassingly parallel" nature of many steps in the formation of the w-stacks, with high performance FFT's performed by the CUFFT library.

A comparison of the execution times for the CUDA and CPU degridders are shown in Figure 6.19. In Figure 6.20 the ratio of the execution time of the w-stacking to the convolution is shown. As the number of visibilities increase, the convolution becomes the dominant cost, following our intuition developed so far in this thesis.

We notice that the unoptimised deconvolution on the GPU is exceptionally slow, showing the pitfalls of unintelligently adapting algorithms for the GPU and SIMT paradigm. However, with the usage of the optimised degridding kernel supporting $8(u) \times 8(v) \times 4(w)$ convolutions we can degrid 15×10^6 visibilities in less than 200ms. In conjunction with the excellent w-stacking performance, we are able to present a highly efficient implementation of the Improved w-stacking algorithm which is science ready.

To understand the impact of convolutional oversampling, which proves to be critical for accuracy, the benchmarking was re-ran with convolution kernels oversampling at a much higher rate. These results are seen in Table 6.9. The optimised cuda code is not particularly effected, with runtime performance degraded for the unoptimised CPU and CUDA code.

The CUDA code has excellent performance, surprisingly so, almost all from optimisations to the memory access patterns of the kernel, so as to maximise the memory bandwidth of the memory bus, as well using it efficiently. For the parameters in Table 6.7 , we require approximately 62 GB of memory to be loaded. The P100 has a stated memory bandwidth of

Configuration	*w*-stacking Time (ms)	Convolution Time (ms)
CUDA	89	146
CUDA (unoptimised)	89	796
CPU	1650	281

Table 6.8: Execution times for Improved *w*-stacking code.

Configuration	*w*-stacking Time (ms)	Convolution Time (ms)
CUDA	89	153
CUDA (unoptimised)	89	1282
CPU	1567	330

Table 6.9: Execution Times for Improved *w*-stacking with high oversampling resolution convolution kernels. High oversampling constitutes an oversampling rate of 65536 per support point, compared to 4096 for the low oversampling regime.

549 GB/s, which should allow all data to be read from memory, assuming 100% efficiency, in 0.113 seconds. Thus for our current implementation we are running at approximately 80% efficiency in terms of peak memory bandwidth. The last 20% of memory bandwidth efficiency may be realised through manual pre-fetching.

Additional optimisation could be performed by reusing grid and kernel accesses wherever possible to reduce the necessary memory requirement. In the general case (of almost randomly distributed) visibilities this will likely be very difficult.

6.3.4.2 Accuracy

Table 6.10 shows the average accuracy for two different rates of oversampling. The low rate oversampling constitutes oversampling the optimal grid convolutional function at a rate of 4096. The gridding correction is sampled at a rate of $1 + 1 \times 10^4$. For high rate oversampling, the oversampling is increased to a rate of 65536 per support point for the grid convolution function, and the gridding correction is sampled at a rate of $1 + 1 \times 10^5$.

It emerged that the absolute accuracy provided by the optimal convolution functions is significantly better than that available at a practical oversampling level. Interpolation was not considered in this analysis, but it ought to be considered in the continuing development of this scheme, particularly in extensions toward imaging and the "gridding" problem. GPU's are often equipped with texture memory which allows high speed hardware interpolation at the cost of some extra memory management. The accuracy figures reported here are broadly consistent with the oversampling analysis in Chapter 3 of Ye (2019) (see in particular Figure 3.40 of that work).

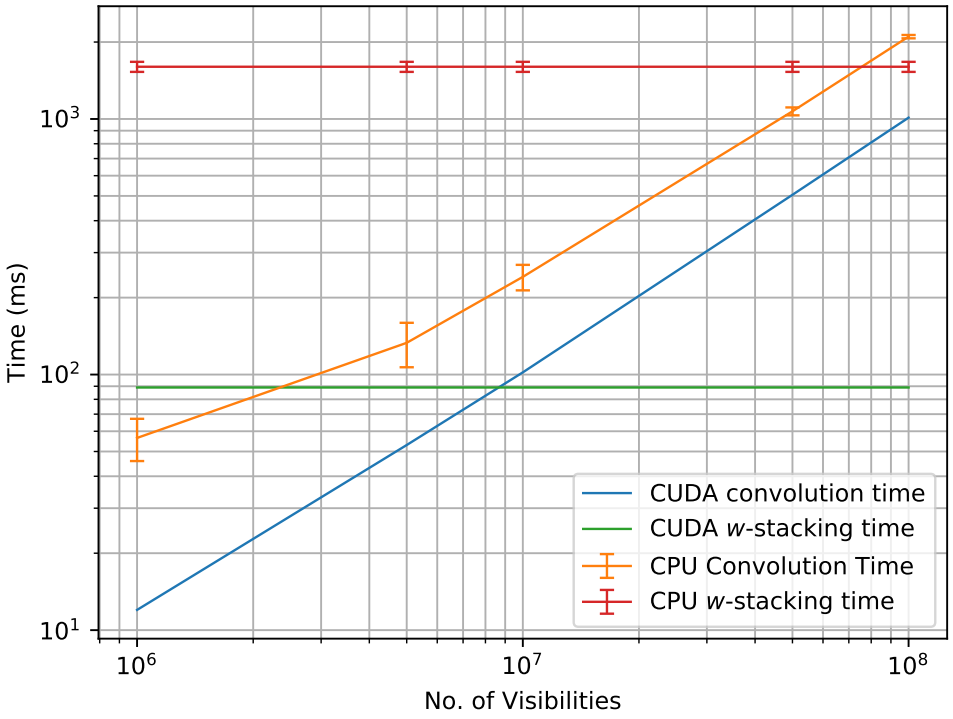

Figure 6.19: Execution time versus number of visibilities for the Optimised CUDA and CPU Improved *w*-stacking code. Error bars are not shown for the CUDA code due to the extremely small standard deviation in runtime. For each data point, the algorithm was ran 10 times.

Oversampling	Mean Absolute Error
Low	1e-3
High	2e-4

Table 6.10: Average accuracy for each convolution scheme. This is the Mean Absolute Error between a visibility and the visibility calculated with an exact DFT calculation.

6.4 Discussion

This chapter has presented work done toward a broad and ambitious scheme of improving and optimising wide-field imaging to try and satisfy the need for high accuracy as well as fast gridding and degridding algorithms for planned and future interferometric arrays. Current schemes are not feasible for an instrument of the size and scale of the Square Kilometer Array, and as it currently stands the results presented represent a definitive step forward.

The original *w*-towers framework offers something that is not often seen in wide-field imaging techniques: an ability to make the problem highly parallel. That is, it removes the need

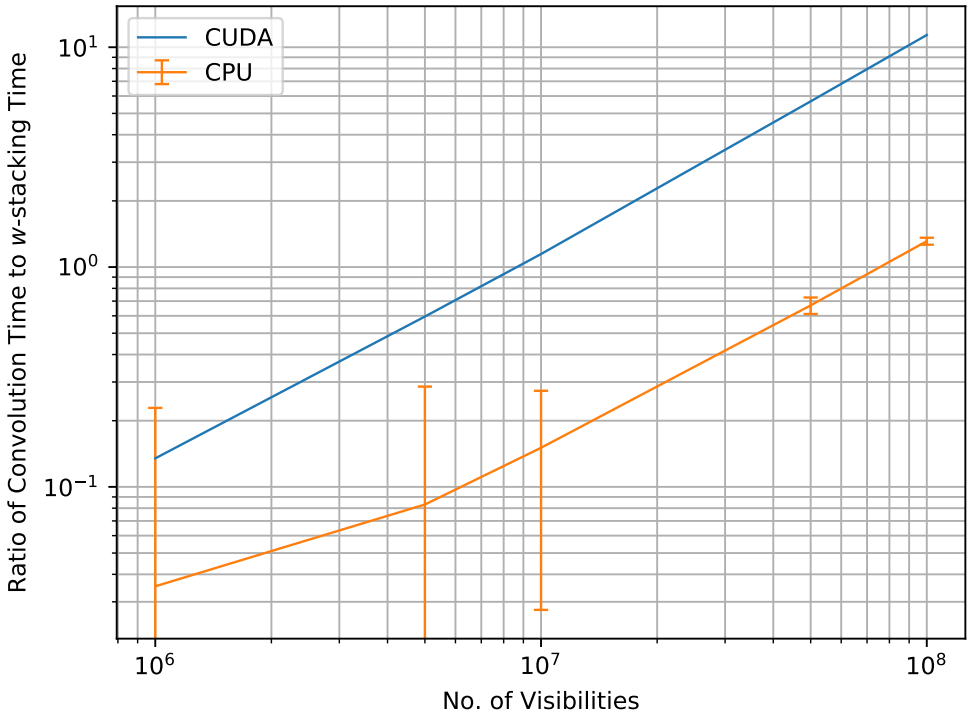

Figure 6.20: Ratio between *w*-stacking element and convolution, plotted versus the number of visibilities for the Optimised CUDA and CPU Improved *w*-stacking code. Beyond a certain number of visibilities the convolution becomes the dominant cost in the execution of the Improved *w*-stacking algorithm. Error bars are not shown for the CUDA code due to the extremely small standard deviation in runtime. For each data point, the algorithm was ran 10 times.

for explicit synchronisation constructs within the algorithm. It offers accuracy comparable to incumbent techniques with relatively strong performance.

As we have discovered, however, there is a flaw in partitioning the grid into sub-grids. This flaw is due to the inhomogeneity of workload distributed across the u, v plane in the number of visibilities that exist. Many extant radio interferometers have a relatively dense central core, with further flung antennas to form longer baselines, facilitating greater angular resolution. This of course creates large numbers of short baselines that are located close to the origin in the u, v coordinates of the spatial coherence vector space. As a result, whilst processing the outlying towers is efficient and takes very little time, the central towers take longer to process due to the necessity to grid many visibilities close to the origin, meaning that the runtime of the *w*-towers algorithm is ultimately constrained by the central towers. There is also the need to bin the visibilities for each tower, a very time consuming process.

Additional schemes to partition the grid, including hierarchical gridding and others, were not considered within this work, because of the development of the Improved *w*-towers framework for arbitrary placement of the towers and the ability to image or predict visibilities across an entire cluster. This is a remarkable innovation, and is likely to be the only way to image SKA-scale datasets at the accuracy and speed required.

Moving on from consideration of the original *w*-towers framework, focus was placed on a new method of aperture synthesis based on the shining innovations in Tan (1986), and expanded with the usage of the Improved *w*-stacking framework in Ye (2019). An extremely efficient and highly parallelised Improved *w*-stacking degridder is presented, using the Optimal Convolution Functions. This has excellent accuracy, comparable with Image Domain Gridding, but crucially the ability to be almost arbitrarily accurate with the use of the correctly optimised Grid Convolution Function and Grid Correction Function. It has been demonstrated that this scheme can be efficiently implemented on modern GPU hardware.

Optimisation was performed in the knowledge that the correct way to optimise the algorithms normally found in aperture synthesis is not to concentrate on computing optimisations, but rather to optimise the memory access patterns to most efficiently take advantage of the memory bandwidth available. Indeed, all of the optimisations and performance increases are due to these these memory-aware techniques. As shown, we can utilise practically 100% of the available memory bandwidth on a modern GPU, and be highly efficient on a CPU by taking into account effects such as cache thrashing as well as cache locality and efficient loads and writes on the memory bus.

Plenty more work lies ahead to improve this technique further. It is proposed to extend this work so as to include a highly optimised implementation of imaging using the Improved *w*-stacking framework within the context of the Improved *w*-towers framework, with the eventual merger of the two schemes into a single software package. The imaging problem presents greater difficulties in assessing accuracy of pixels (although this can be done using the predictive routines described here), as well as the greater complexity of efficiently implementing gridding versus degridding. Gridding will require atomic updates to the grid in order to ensure there are no race conditions on the 3-D *w*-stack pixels.

Going forwards, we think this presents an opportunity for truly perfect wide-field imaging with visibilities in a way that is amenable to high performance computing techniques with next generation telescopes. As was shown in Chapter 3, a seemingly intractable problem can become very tractable with the right techniques, with previously unseen benefits. The schemes described here can, with further development, satisfy all of the requirements of these next generation telescopes.

Chapter 7

Summary and Closing Remarks

This thesis has presented several new methods that improve upon the state of the art in interferometric imaging. These advances have been made by questioning the underlying assumptions and mathematics of radio interferometry so as to yield some surprising results.

In Chapter 2, the first working direct imaging correlator is presented. This builds on decades of advances in convolutional gridding and underlying knowledge of fourier theory, as well as parallel advancements in computational hardware. It is, to the best of the author's knowledge, the only direct imaging correlator ever demonstrated that is completely generic in the array's that it can be applied to. Indeed, the only major constraint is bandwidth, which is being improved upon continually. Perfecting the algorithms through continued optimisation, alongside the current growth in the capabilities of massively parallel hardware such as GPU's, will make this scheme more capable at higher frequencies.

Within Chapter 3, there was a further innovation on the advance made in Chapter 2 by again interrogating underlying assumptions about the mathematics of interferometry, and how these mathematical relations are themselves related to the compute hardware on which they are implemented. It was shown that current wide-field imaging techniques are not feasible for a direct imaging correlator, due to the hard real-time requirements in place. A solution to the wide-field problem through a direct discrete fourier transform modeled as a linear transform was presented, which is extremely efficient when implemented on GPU's due to them being well suited for a compute bound algorithm such as a matrix multiplication.

This provides an advance in the state of the art in interferometric imaging which had been previously thought impossible. However, there is still much work to be done in developing

this capability further. Calibration and deconvolution remain open problems, with real-time calibration presenting unique challenges, especially in the area of RFI mitigation. This difficulty has been considered in Chapters 2 and 3, with discussion as to how calibration schemes can be implemented. Deconvolution of the very large numbers of images produced by this scheme, especially at high time resolution, is a challenging problem. Further research will be needed to be done in creating a deconvolution framework that can deal with this unique requirement.

Moving on to the use of higher order data products for scientific analysis, in Chapter 4 we use the closure phase spectrum of the radio interferometer, the bispectrum, as a method for analysing discrepancies in the design of HERA, an interferometer with a redundant layout. The unique properties of the closure phase make it immune to direction independent gains resulting from individual antennas and calibration. This lack of requirement for calibration allows it to be a particularly interesting tool for the probing of non-redundancies in a supposedly redundant array. The scheme was used together with forward modeling of a HERA-like array to understand and constrain the main drivers of non-redundancies in the closure phase. This will provide a valuable troubleshooting tool for diagnosing these redundant arrays going forwards.

The original reason for this redundancy analysis, was to understand these non-redundancies relative to using the bispectrum for a cosmological measurement of the HI line during the Epoch of Reionisation. In Chapter 5, a cosmological limit on the EoR signal was derived using the bispectrum approach. The conclusion is that currently we are data limited; the cosmological signal is overpowered by both thermal noise and systematics from the telescope. The reduction of thermal noise will require a much larger amount of data to be collected and integrated over to reduce the noise temperature to below the cosmological temperature. How to remove systematic effects in the telescope without also removing cosmological signals is currently unknown; systematics include reflections and non-redundancies in the beam shapes and pointings. This is a major open problem. A discussion is also presented of the effects of noise behavior when using a higher order data product such as the bispectrum, and a thorough supplementary analysis undertaken in Appendix A.

There are several problems that remain open within this scheme. One of the most interesting, and thus far unanswered, is how very small signals such as the HI line, are affected by non-redundancies and systematic effects in a telescope. Within this, an understanding of how calibration affects the signal within the presence of non-redundancies and systematics is also required. With the expansion of the HERA array and the current exertions underway to constrain and fix systematic and non-redundancy errors, there is a need to expand the analysis presented here to this new dataset. Larger amounts of improved data will help drive down the thermal noise floor as currently we are mostly thermal noise limited in addition to the systematics.

Finally in Chapter 6, moving back to the problem of improving the state of the art in

interferometric imaging, the problem of wide-field imaging with visibilities was examined in the context of future state of the art arrays such as the Square Kilometer Array, and next generation Very Large Array. A novel imaging scheme, *w*-towers, is presented, that has comparable performance to *w*-projection but with increased accuracy and performance benefits when run on a CPU. This scheme is superseded by an improved version of *w*-towers that allows the problem to be distributed across a compute cluster as a set of sub-grids and sub-facets to process. This processing was performed using Improved *w*-stacking with optimal convolution functions, with a heavily optimised GPU implementation.

There is much work that remains to be done within this scheme, such as producing an optimised implementation of the Improved *w*-stacker for imaging, and integration of these optimal implementations with the Improved *w*-towers recombination framework to facilitate a distributed wide-field imager with the accuracy and speed required for imaging next generation arrays.

Appendix A

Interferometric Noise Properties

The properties of the noise on individual antennas, and the resulting visibilities and higher order data products, are important components of radio interferometric imaging and analysis. There is a deep link between the behaviour of the system, including the antennas and the overall design, and the errors and biases in its data products as it looks at the sky.

This appendix derives rigorously and describes the noise in a radio interferometer, and the consequent effects on two mathematical quantities: the baseline phasor, also known as a 'visibility' (Thompson et al. 2017), and the Closure Phase (Jennison 1958). This section is heavily influenced by Kulkarni (1989), and sets out to rigorously describe the mathematical behaviour of the radio interferometer through a direct, brute force expansion of the statistical moments of the visibility and the closure phase.

A.1 The Visibility Noise

A visibility is defined as the cross-correlation operation of two voltages from antennas a and b:

$$V_{ab} = \int_{-\infty}^{\infty} v_a(t) v_b(t+\tau) dt \tag{A.1}$$

where $v_{a/b}$ represents the voltage at each antenna, t is time, and τ the delay of the cross-correlation operation. Each v term consists of a signal term and a complex noise term such that $v_a = v_a^S + v_a^N$, with superscripts S and N indexing the signal and noise terms, respectively.

Most modern radio inteferometers are digital in nature and thus V_{ab} is defined on a set of discrete basis functions representing the discrete sampling of the analogue to digital converters

on the interferometer. Thus V_{ab} becomes:

$$V_{ab} = \sum_{t=-\infty}^{\infty} v_a[t] v_b[t+\tau] \tag{A.2}$$

The Wiener-Khinchin theorem allows us to re-arrange Equation A.1 as a multiplication with the use of a fourier transform of the voltages:

$$E_a(\nu) = \int_{\infty}^{\infty} v_a(t) e^{-i2\pi t\nu} dt \tag{A.3}$$

and with the corollary quantity for a discrete set of basis functions:

$$E_a(\nu) = \sum_{-\infty}^{\infty} v_a[t] e^{-i2\pi t\nu} \tag{A.4}$$

with $E_a = E_a^S + E_a^N$ as with the voltages. Thus via the Wiener-Khinchin theorem:

$$V_{ab} = E_a E_b \tag{A.5}$$

$$V_{ab} = \left[E_a^S + E_a^N\right] \cdot \left[E_b^S + E_b^N\right] \tag{A.6}$$

It is customary to average V_{ab} over some N samples where N is often large, with the average corresponding to the arithmetic pythagorean mean. We will denote this average by $\langle\rangle$:

$$V_{ab} = \langle E_a E_b \rangle \tag{A.7}$$

$$V_{ab} = \left\langle \left[E_a^S + E_a^N\right] \cdot \left[E_b^S + E_b^N\right] \right\rangle \tag{A.8}$$

Expanding out Equation A.7, we end up with:

$$V_{ab} = \left\langle E_a^S E_b^{S*} + E_a^S E_b^{N*} + E_b^{S*} E_a^N + E_a^N E_b^N \right\rangle \tag{A.9}$$

$$V_{ab} = \left\langle E_a^S E_b^{S*} \right\rangle + \left\langle E_a^S E_b^{*N} \right\rangle + \left\langle E_b^{S*} E_a^N \right\rangle + \left\langle E_a^N E_b^{*N} \right\rangle \tag{A.10}$$

The statistical properties of the noise are such that noise on different antennas should be independent. It is assumed that the real and imaginary components of v_a^N are independent and identically distributed, and thus uncorrelated:

$$\left\langle \Re(v_a^N) \Im(v_a^N) \right\rangle \overset{\text{i.i.d.}}{=} 0 \tag{A.11}$$

It will be proven later that noise on a baseline is also independent. Thus:

$$\left\langle E_a^S E_b^{S*} \right\rangle \neq 0 \tag{A.12}$$

$$\left\langle E_a^S E_b^{*N} \right\rangle = \left\langle E_b^{S*} E_a^N \right\rangle = \left\langle E_a^N E_b^{*N} \right\rangle = 0 \tag{A.13}$$

The reason that $\left\langle E_a^S E_b^{S*} \right\rangle \neq 0$ is because of the correlated sky signal that they share. Noise terms will not correlate with the sky signal or with each other due to statistical independence. We can simplify V_{ab} to:

$$V_{ab} = \left\langle V_{ab}^S + V_{ab}^N \right\rangle \tag{A.14}$$

where $V_{ab}^S = \left\langle E_a^S E_b^{S*} \right\rangle$, and $V_{ab}^N = \left\langle E_a^S E_b^{*N} \right\rangle + \left\langle E_b^{S*} E_a^N \right\rangle + \left\langle E_a^N E_b^{*N} \right\rangle$

Using the same method, it can be shown that the autocorrelation of an antenna is wholly real:

$$V_{aa} = \left\langle \left[E_a^S + E_a^N\right] \cdot \left[E_a^S + E_a^N\right] \right\rangle \tag{A.15}$$

$$= \left\langle E_a^S E_a^{S*} + E_a^S E_a^{N*} + E_a^{S*} E_a^N + E_a^N E_a^{N*} \right\rangle \tag{A.16}$$

$$= |E_a^S|^2 + 2\Re\left[E_a^S E_a^{N*}\right] + |E_a^N|^2 \tag{A.17}$$

A.1.1 Closure Phase

The closure phase (Jennison 1958), can be described as the addition of the phasors of three baselines forming a closed triangle from three antennas:

$$\phi_{abc} = \phi_{ab} + \phi_{bc} + \phi_{ca} + \phi_{abc}^N \tag{A.18}$$

where each ϕ component consists of a phasor due to the sky and an antenna-dependent term, such that $\phi_{ab} = \theta_{ab}^s + (\chi_a - \chi_b)$. Thus Equation A.18 becomes:

$$\phi_{abc} = \theta_{ab}^s + (\chi_a - \chi_b) + \theta_{bc}^S + (\chi_b - \chi_c) + \theta_{ca}^S + (\chi_c - \chi_a) + \phi_{abc}^N \tag{A.19}$$

where ϕ_{abc}^N is thermal noise. The cancellation of the antenna-dependent terms is obvious and well known and utilised in various areas of radio interferometry. Another method of defining Equation A.18 is:

$$C_{abc} = V_{ab} V_{bc} V_{ca} \tag{A.20}$$

$$= |V_{ab}||V_{bc}||V_{ca}| \exp\left[i(\phi_{ab} + \phi_{bc} + \phi_{ca})\right] \tag{A.21}$$

$$= |V_{ab}^S + V_{ab}^N||V_{bc}^S + V_{bc}^N||V_{ca}^S + V_{ca}^N| \exp\left[i(\phi_{ab} + \phi_{bc} + \phi_{ca} + \phi_{abc}^N)\right] \tag{A.22}$$

where the last two equations are equivalent in the limit of an infinite number of measurements, as the noise term will gradually reduce to zero by the averaging process.

As can be seen, Equation A.18 can be derived from Equation A.20 with the relationship $\phi_{abc} = arg\left[C_{abc}\right]$

A.1.2 Gaussianity

One of the major assumptions in radio inteferometry is that the noise on the voltages is Gaussian. If v_a^N is non-Gaussian then this could complicate our analysis, especially as the Gaussian is an eigenfunction of the fourier transform, and thus it will make analysis easier if gaussianity is true. Fortunately, due to the visibilities being averaged many times, the Central Limit Theorem (Feller 1945) can be invoked, and we can move on with the assumption of Gaussian noise in the asymptotic limit of many measurements.

A.2 Signal to Noise Ratio

From above, we can define $E_a E_b^*$ as our 'signal', and all other terms as noise. From this the signal-to-noise ratio (SNR) can be defined as :

$$SNR = \frac{\left\langle E_a^S E_b^{S*} \right\rangle_N}{\left\langle E_a^S E_b^{*N} + E_b^{S*} E_a^N + E_a^N E_b^{*N} \right\rangle_N} \tag{A.23}$$

where $\left\langle \dots \right\rangle_N$ indicates the average over N measurements.

The importance of the SNR in interferometry can be realised through plotting the vectors of the different components of a visibility. This is shown in Figure A.1, which gives a visual reference of how the signal and noise is related to the distribution of the phase.

A.3 Distribution of Visibility Phasor

V_{ab} can be described as $\Re(V_{ab}) + i\Im(V_{ab}) = |V_{ab}|e^{i\phi_{ab}}$. Thompson et al. (2017) offer derived distributions for this in the regime where $|V_{ab}^S| >> |V_{ab}^N|$:

$$P(|V_{ab}|) = \frac{|V_{ab}|}{\sigma_{ab}^2} \exp\left[-\frac{|V_{ab}|^2 + |V_{ab}^S|^2}{2\sigma_{ab}^2} \right] I_0\left[\frac{|V_{ab}|^2 |V_{ab}^S|}{\sigma_{ab}^2} \right] \tag{A.24}$$

$$P(\phi_{ab}) = \frac{1}{2\pi} \exp\left[\frac{-|V_{ab}^S|^2}{2\sigma^2} \right] \left\{ 1 + \sqrt{\frac{\pi}{2}} \frac{|V_{ab}^S| \cos(\phi_{ab})}{\sigma_{ab}} \times \right. \tag{A.25}$$

$$\left. \exp\left[\frac{|V_{ab}^S|^2 \cos^2 \phi_{ab}}{2\sigma_{ab}^2} \right] \left[1 + erf\left(\frac{|V_{ab}^S| cos\phi_{ab}}{\sqrt{2}\sigma_{ab}} \right) \right] \right\} \tag{A.26}$$

where $erf(x)$ is the error function:

$$erf(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\, dt \tag{A.27}$$

$$erf\left(\frac{x}{\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\frac{t^2}{2}}\, dt \tag{A.28}$$

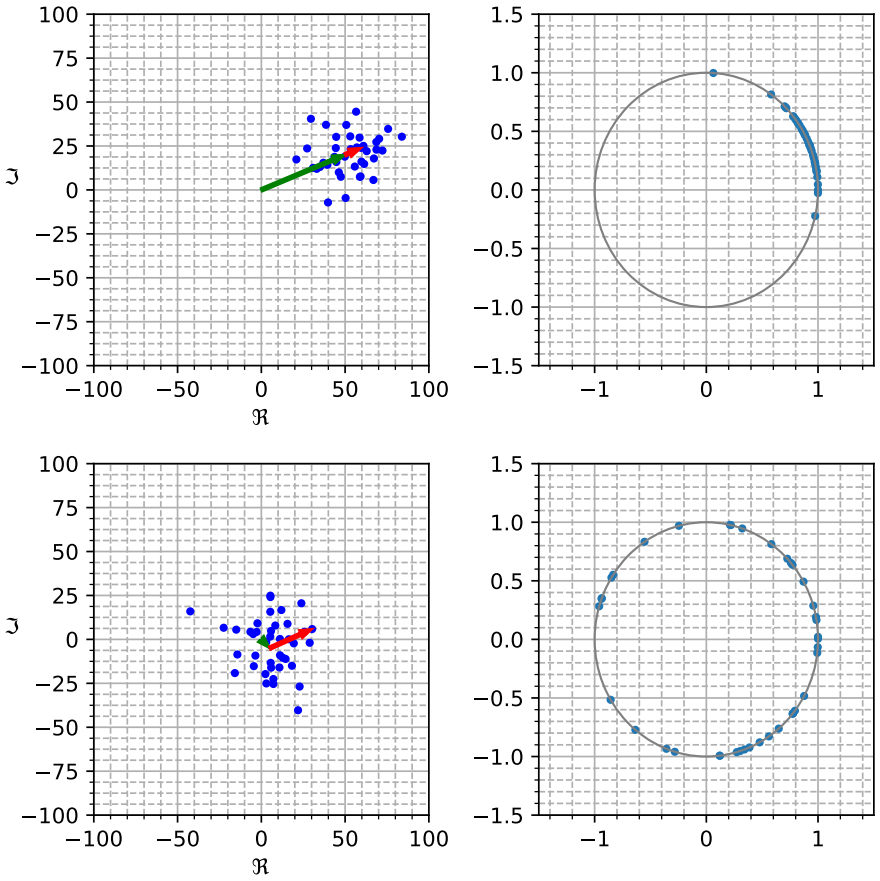

Figure A.1: A graphical aid for viewing the relationship between the signal to noise ratio and the distribution of the phase around the unit circle. The green arrow corresponds to the correlated sky signal, and the red arrow to a noise vector added to this sky signal. The right two plots show the distribution of the phase of each simulated phasor around the unit circle. In the limit of a low signal to noise ratio, the distribution of the phase approaches a uniform circular distribution.

The analytical representation of $|V_{ab}|$ is relatively simple, however ϕ_{ab} is less so. A simpler *prima facie* interpretation of the distribution of the phase follows upon using the Wrapped Gaussian Distribution(also known as the Wrapped Normal Disitribution (WN)) and von Mises Distribution(vMD), of which the characteristic functions are:

$$\int_{2\pi} e^{in\theta} f_{wn}(\theta;\mu,\sigma)d\theta = \exp(in\mu - \frac{n^2\sigma^2}{2}) \tag{A.29}$$

$$f_{vMD}(\theta;\mu,k) = \frac{1}{2\pi}\left[1 + 2\sum_{n=1}^{\infty}\cos(n\theta)\frac{I_{|n|}(k)}{I_0(k)}\right] \tag{A.30}$$

where $I_n(k)$ represents the modified Bessel function of the first kind. The von Mises distribution and the Wrapped Gaussian distribution closely approximate one another. The von Mises

distribution is often easier to work with due to its simpler entropy expression and moment expression, however when working from the data the wrapped Gaussian provides much greater simplicity in estimating sample statistical moments. In both cases, as $k \rightarrow 0$ for the vMD, and as $\sigma \rightarrow \infty$, they both asymptotically approach a uniform circular distribution.

A.3.1 Wrapped Gaussian Statistical Moments

Considering integer values of n in Equation A.29, we end up with:

$$\exp\left[i\bar{\theta} - \frac{\sigma^2}{2}\right] = \bar{R}\exp\left[i\bar{\phi}\right] \tag{A.31}$$

where $\bar{\phi}$ represents the average angular value, and $\bar{R}$ the mean resultant vector, such that:

$$\bar{R}\exp\left[i\bar{\theta}\right] = \frac{1}{N}\sum_{i=0}^{N}\cos\phi_i + \frac{i}{N}\sum_{i=0}^{N}\sin\phi_i \tag{A.32}$$

therefore:

$$\bar{\phi} = \tan^{-1}\left(\frac{\frac{1}{N}\sum_{i=0}^{N}\cos\phi_i}{\frac{1}{N}\sum_{i=0}^{N}\sin\phi_i}\right) \tag{A.33}$$

Equation A.31 can be re-arranged to derive the standard deviation of the wrapped Gaussian:

$$\sigma = \sqrt{-2\ln(R)} \tag{A.34}$$

A.3.2 Observations on Statistical Behaviour of Phasor

The following inferences may be drawn from these derivations:

1) For the limit $SNR \rightarrow 0$ the phasor on a baseline becomes probabilistically a uniform circular distribution because $\bar{R} \rightarrow 0$.
2) For $SNR >> 0$, the phasor on a baseline can be described by a wrapped Gaussian distribution or von Mises Distribution.
3) A corollary is that if the SNR of any baseline in a closure triad C_{abc} approaches zero, then the the probability distribution of the phase of the closure triad, ϕ_{abc} becomes a uniform circular distribution. This assists in identifying malfunctioning antennas in an interferometric array.

A.4 Variance Analysis

In this section the variances and covariances of the visibility and the closure phase will be derived in the same method as Kulkarni (1989). However in our analysis we will make the

assumption that all noise terms are independent of each other and of the correlated signal on each visibility phasor. The variance of a complex random variable can be computed using the pseudovariance:

$$\sigma_{12}^2 = \langle a_1 a_2^* \rangle - \langle a_1 \rangle \langle a_2^* \rangle \tag{A.35}$$

In this analysis it is worthwhile to note the use of the fourth order moment theory for Gaussian random variables:

$$\langle x_1 x_2 x_3 x_4 \rangle = \langle x_1 x_2 \rangle \langle x_3 x_4 \rangle + \langle x_1 x_3 \rangle \langle x_2 x_4 \rangle + \langle x_1 x_4 \rangle \langle x_2 x_3 \rangle \tag{A.36}$$

$$\langle z_1 z_2^* z_3 z_4^* \rangle = \langle z_1 z_2^* \rangle \langle z_3 z_4^* \rangle + \langle z_1 z_3 \rangle \langle z_2^* z_4^* \rangle + \langle z_1 z_4^* \rangle \langle z_2^* z_3 \rangle \tag{A.37}$$

where x is a real Gaussian variable and z is a complex Gaussian variable. This is in itself a limiting case of the more general Isserlis Theorem (which is derived by generalising the fourth order case), a well known probability theorem which is exploited in various Quantum Field Theories in the form of Wick's Theorem:

$$\langle x_1 x_2 \dots x_n \rangle = \sum_{p \in P_n^2} \prod_{\{i,j\} \in p} \langle x_i x_j \rangle \tag{A.38}$$

Appendix A in Kulkarni (1989) provides a proof that Equations A.36 and A.37 are the same for both real and complex Gaussian variables. A summary of the assumptions in the following analysis follows:

1) $\langle z_1 z_2 \rangle \langle z_3^* z_4^* \rangle = 0$; Non-conjugated terms do not correlate.
2) $\langle z_1 z_2^* \rangle \langle z_3 z_4^* \rangle \neq 0$. Conjugated terms *might* correlate depending on if they are independent or not.
3) All noise terms are independent of each other. The corollary of this is that the off-diagonal elements of the covariance matrix of $z_1^N, z_2^N, z_3^N, z_4^N$ are zero.

A.4.1 Variance of a Visibility

Define the pseudovariance for $V_{ab} = V_{ab}^S + V_{ab}^N$:

$$\begin{aligned}
\sigma_{V_{ab}}^2 &= \langle (V_{ab}^S + V_{ab}^N)(V_{ab}^{S*} + V_{ab}^{N*}) \rangle - \langle V_{ab} \rangle \langle V_{ab}^* \rangle \\
&= \langle (V_{ab}^S + V_{ab}^N)(V_{ab}^{S*} + V_{ab}^{N*}) \rangle - |V_{ab}|^2 \\
&= \langle |V_{ab}^S|^2 + |V_{ab}^N|^2 + 2 V_{ab}^N V^{S*} \rangle - |V_{ab}^S|^2 \\
&= |V_{ab}^N|^2
\end{aligned} \tag{A.39}$$

If the noise is uncorrelated between time measurements then the variance $\sigma_{V_{ab}}^2 \rightarrow 0$ as the number of independent of measurements $N \rightarrow \infty$.

A.4.2 Covariance of Visibility with another Visibility Sharing One Antenna

Define the pseudo-covariance for $V_{ab} = V^S_{ab} + V^N_{ab}$ and $V_{bc} = V^S_{bc} + V^N_{bc}$

$$\begin{aligned}\sigma^2_{V_{ab},V_{bc}} &= \left\langle (V^S_{ab} + V^N_{ab})(V^{S*}_{bc} + V^{N*}_{bc}) \right\rangle - \left\langle V_{ab} \right\rangle \left\langle V^*_{bc} \right\rangle \\ &= \left\langle (V^S_{ab} + V^N_{ab})(V^{S*}_{ab} + V^{N*}_{ab}) \right\rangle - V^S_{ab} V^{S*}_{bc} \\ &= \left\langle V^S_{ab} V^{S*}_{bc} + V^N_{ab} V^{N*}_{bc} + 2V^S_{ab} V^{N*}_{bc} \right\rangle - V^S_{ab} V^{S*}_{bc} \\ &= 0 \end{aligned} \tag{A.40}$$

Thus two baselines can be considered statistically independent.

A.4.3 Variance of a Closure Triad

To derive the variance of a closure triad we will use Equation A.35 alongside:

$$C_{abc} = (V^S_{ab} + V^N_{ab})(V^S_{bc} + V^N_{bc})(V^S_{ca} + V^N_{ca}) \tag{A.41}$$

Thus we can define the pseudovariance for C_{abc}:

$$\begin{aligned}\sigma^2_{C_{abc}} &= \left\langle C_{abc} C^*_{abc} \right\rangle - \left\langle C_{abc} \right\rangle \left\langle C^*_{abc} \right\rangle \\ &= \left\langle (V^S_{ab} + V^N_{ab})(V^S_{bc} + V^N_{bc})(V^S_{ca} + V^N_{ca}) \times \right. \\ &\quad \left. (V^{S*}_{ab} + V^{N*}_{ab})(V^{S*}_{bc} + V^{N*}_{bc})(V^{S*}_{ca} + V^{N*}_{ca}) \right\rangle - |C_{abc}|^2 \\ &= \left\langle \left[V^S_{ab} V^S_{bc} V^S_{ca} + V^S_{ab} V^S_{bc} V^N_{ca} + V^S_{ab} V^N_{bc} V^S_{ca} + V^S_{ab} V^N_{bc} V^N_{ca} + \right. \right. \\ &\quad V^N_{ab} + V^S_{bc} V^S_{ca} + V^N_{ab} V^S_{bc} V^N_{ca} + V^N_{ab} V^N_{bc} V^S_{ca} + V^N_{ab} V^N_{bc} V^N_{ca} \left. \right] \times \\ &\quad \left[V^{S*}_{ab} V^{S*}_{bc} V^{S*}_{ca} + V^{S*}_{ab} V^{S*}_{bc} V^{N*}_{ca} + V^{S*}_{ab} V^{N*}_{bc} V^{S*}_{ca} + V^{S*}_{ab} V^{N*}_{bc} V^{N*}_{ca} + \right. \\ &\quad \left. \left. V^{N*}_{ab} V^{S*}_{bc} V^{S*}_{ca} + V^{N*}_{ab} V^{S*}_{bc} V^{N*}_{ca} + V^{N*}_{ab} V^{N*}_{bc} V^{S*}_{ca} + V^{N*}_{ab} V^{N*}_{bc} V^{N*}_{ca} \right] \right\rangle \end{aligned} \tag{A.42}$$

After a tedious expansion, which we omit, we find 64 sixth-order terms, of which eight are non-zero, which is a similar answer as Kulkarni (1989). These terms are:

$$A = V^S_{ab} V^S_{bc} V^S_{ca} V^{S*}_{ab} V^{S*}_{bc} V^{S*}_{ca} = C_{abc} C^*_{abc} = |C_{abc}|^2 \tag{A.43}$$

$$B = V^S_{ab} V^S_{bc} V^N_{ca} V^{S*}_{ab} V^{S*}_{bc} V^{N*}_{ca} = |V^S_{ab}|^2 |V^S_{bc}|^2 |V^N_{ca}|^2 \tag{A.44}$$

$$C = V^S_{ab} V^N_{bc} V^S_{ca} V^{S*}_{ab} V^{N*}_{bc} V^{S*}_{ca} = |V^S_{ab}|^2 |V^S_{ca}|^2 |V^N_{bc}|^2 \tag{A.45}$$

$$D = V^N_{ab} V^{S*}_{bc} V^{S*}_{ca} V^{N*}_{ab} V^{S*}_{bc} V^{S*}_{ca} = |V^S_{bc}|^2 |V^S_{ca}|^2 |V^N_{ab}|^2 \tag{A.46}$$

$$E = V^S_{ab} V^N_{bc} V^N_{ca} V^{S*}_{ab} V^{N*}_{bc} V^{N*}_{ca} = |V^S_{ab}|^2 |V^N_{bc}|^2 |V^N_{ca}|^2 \tag{A.47}$$

$$F = V^N_{ab} V^S_{bc} V^N_{ca} V^{N*}_{ab} V^{S*}_{bc} V^{N*}_{ca} = |V^S_{bc}|^2 |V^N_{ab}|^2 |V^N_{ca}|^2 \tag{A.48}$$

$$G = V^N_{ab} V^N_{bc} V^S_{ca} V^{N*}_{ab} V^{N*}_{bc} V^{S*}_{ca} = |V^S_{ca}|^2 |V^N_{ab}|^2 |V^N_{bc}|^2 \tag{A.49}$$

$$H = V^N_{ab} V^N_{bc} V^N_{ca} V^{N*}_{ab} V^{N*}_{bc} V^{N*}_{ca} = |V^N_{ab}|^2 |V^N_{bc}|^2 |V^N_{ca}|^2 \tag{A.50}$$

If our assumption that the noises are not independent from each other fail, so that there is a covariance between each noise term, then this analysis becomes much more complicated, with the result that variance will increase. With the above in mind, our covariance becomes:

$$\sigma^2_{C_{abc}} = \frac{1}{N}\sum_{i=0}^{N}(A_i + B_i + C_i + D_i + E_i + F_i + G_i + H_i - |C_{abc}|^2) \tag{A.51}$$

$$\sigma_{C_{abc}} = \frac{1}{\sqrt{N}}\sum_{i=0}^{N}(B_i + C_i + D_i + E_i + F_i + G_i + H_i)^{\frac{1}{2}} \tag{A.52}$$

A.4.4 Covariance of Two Triads with a Shared Antenna

Firstly, define the covariance between two closure triads sharing a common antenna:

$$\sigma^2_{C_{abc},C_{cde}} = \left\langle C_{abc}C^*_{cde}\right\rangle - \left\langle C_{abc}\right\rangle\left\langle C^*_{cde}\right\rangle \tag{A.53}$$

One term survives the statistical averaging process:

$$A = C_{abc}C^*_{cde} \tag{A.54}$$

Which results in:

$$\sigma^2_{C_{abc},C_{cde}} = 0 \tag{A.55}$$

in the limit of $N \rightarrow \infty$. Thus we can state that two closure triads which share an antenna are statistically independent, provided that all noise terms are independent from each other. In the case of $N < \infty$ the covariance will reduce as $\frac{1}{\sqrt{N}}$. It is also worthwhile to note that the statistical fluctuations of the sky signal are also independent.

A.4.5 Covariance of Two Triads with a Shared Baseline

Firstly, define the covariance between two closure triads sharing a common baseline:

$$\sigma^2_{C_{abc},C_{abd}} = \left\langle C_{abc}C^*_{abd}\right\rangle - /big < C_{abc}\left\rangle\left\langle C^*_{abd}\right\rangle \tag{A.56}$$

Two terms survive statistical averaging:

$$A = V^S_{ab}V^S_{bc}V^S_{ca}V^{S*}_{ab}V^{S*}_{bd}V^{S*}_{da} = C_{abc}C^*_{abd} \tag{A.57}$$

$$B = V^N_{ab}V^S_{bc}V^S_{ca}V^{N*}_{ab}V^{S*}_{bd}V^{S*}_{da} \tag{A.58}$$

The term B can be interpreted as a phasor:

$$B = |V^N_{ab}|^2|V^S_{ca}||V^S_{bc}|V^S_{de}||V^S_{bd}|\exp\left[i(\phi_{ca} + \phi_{bc} - \phi_{de} - \phi_{bd})\right] \tag{A.59}$$

Upon defining $\phi_{abd} = \phi_{ab}+\phi_{bd}+\phi_{da}$ and recognising that $\phi_{abc}-\phi_{abd} = \phi_{bc}+\phi_{ca}-\phi_{bc}-\phi_{de}$, we can see that B is:

$$B = |V^N_{ab}|^2|V^S_{ca}||V^S_{bc}|V^S_{de}||V^S_{bd}| \exp\left[i(\phi_{abc} - \phi_{abd})\right] \tag{A.60}$$

Therefore in the limit of $N \rightarrow \infty$:

$$\sigma^2_{C_{abc},C_{abd}} = |V^N_{ab}|^2|V^S_{ca}||V^S_{bc}|V^S_{de}||V^S_{bd}| \exp\left[i(\phi_{abc} - \phi_{abd})\right] \tag{A.61}$$

Thus two triads which share a baseline are not statistically independent.

Appendix B

Direct Electric Field Imaging in Spherical Harmonics Basis

It has been shown in McEwen and Scaife (2008), and later by Carozzi (2015), that it is possible to expand a visibility distribution in a basis of spherical harmonic functions. For direct imaging, it would be of interest if a similar technique could be used for interferometers with electric fields directly. This would allow the wide-field imaging problem to be solved by casting the van-Cittert Zernike theorem in spherical co-ordinates. Indeed this was one of the main motivations discussed in Carozzi (2015).

Here a solution is derived for a single polarisation, thus simplifying the treatment to that of a scalar electric field. This extends the work in Carozzi (2015) to interferometry done with direct imaging. In Section B.1 the extension of the Spherical Wave Harmonic Transform to electric fields is presented, and Section B.2 discusses the practicalities of this and why it is a dubious proposition to attempt to solve the wide field imaging problem in this way for electric field direct imaging.

B.1 Theory

B.1.1 Electric Field Imaging in Spherical Co-ordinates

Firstly we will start by defining two co-ordinate bases:

- r - Position vector in measurement plane.
- k - Wave vector.

- $\Omega_k = (\theta_k, \phi_k)$ - The angular components of k, the fourier modes on the sky.
- $\Omega_r = (\theta_r, \phi_r)$ - The angular components of r, the antenna locations in a spherical co-ordinate system.

Next, we state the van-Cittert Zernike (vCZ) theorem in these co-ordinate bases:

$$V(r,k) = \int I(\Omega_k) e^{-ik\cdot r} d\Omega_k \tag{B.1}$$

$$I(\Omega_k) = \int V(r,k) e^{ik\cdot r} dr \tag{B.2}$$

Using the multiplication convolution theorem, we can also re-express the intensity distribution of the sky using the electric fields measured at each antenna:

$$I(\Omega_k) = \left[\int E(r,k) e^{ik\cdot r} dr\right] \left[\int E(r,k) e^{ik\cdot r} dr\right]^* \tag{B.3}$$

B.1.2 Deriving Harmonic Co-efficients

From Chapter 9 of Jackson (2007) we can find the general solution of a scalar electric field in spherical co-ordinates:

$$E(r,k) = \sum_{l,m} \left[E^{\alpha}_{lm} h^{\alpha}_l(kr) + E^{\beta}_{lm} h^{\beta}_l b(kr)\right] Y_{lm}(\Omega_r) \tag{B.4}$$

where $h^{\alpha,\beta}_l = j_l(kr) \pm i n_l(kr)$, with j_l representing the spherical Bessel function, and n_l representing the Hankel functions, where α and β represent the labels for the conjugates of h. E^{α} and E^{β} can be assumed to be the same up to a normalisation constant. Y_{lm} represents the spherical harmonic function of order l and degree m. As kr is a real parameter, it follows that h^{α}_l and h^{β}_l are complex conjugates of one another. With this in mind we can reduce the above to:

$$E(r,k) = 2 \sum_{lm} E_{lm} j_l(kr) Y_{lm}(\Omega_r) \tag{B.5}$$

With this completed, we can begin to derive the electric field contributions from the sky to a point in the measurement plane (say an antenna position):

$$E(r,k) = \int E(\Omega_k) e^{-k\cdot r} d\Omega_k \tag{B.6}$$

By applying the Laplacian operator to the above, it is easy to show that this relation satisfies the Helmholtz equation, thus a valid wave equation:

$$\nabla^2 E(r,k) + k^2 E(r,k) = 0 \tag{B.7}$$

Whereas before using the van Cittert-Zernike theorem we deal with a real intensity distribution, we now have a complex sky, assuming that each point on the sky has a random complex phase compared to other points. Next we expand out the sky into spherical harmonics:

$$E(\Omega_k) = \sum_{lm} e_{lm} Y_{lm}(\Omega_k) \tag{B.8}$$

Next we use the plane wave decomposition formula (again see Jackson (2007)):

$$e^{-ik\cdot r} = 4\pi \sum_{lm} (-i)^l j_l(kr) Y_{lm}(\Omega_r) Y^*_{lm}(\Omega_k) \tag{B.9}$$

Thus $E(r,k)$ becomes:

$$E(r,k) = 4\pi \int \sum_{lm} e_{lm} Y_{lm}(\Omega_k)(-i)^l j_l(kr) Y_l m(\Omega_r) Y^*_{lm}(\Omega_k) \tag{B.10}$$

We can use the orthonormality of the spherical harmonic basis functions to simplify Equation B.10:

$$\int_0^{4\pi} Y_{lm}(\Omega_k) Y^*_{l'm'}(\Omega_k) d\Omega = \delta_{ll'}\delta_{mm'} \tag{B.11}$$

Therefore Equation B.10 becomes:

$$E(r,k) = 4\pi \sum_{lm} e_{lm}(-i)^l j_l(kr) Y_{lm}(\Omega_r) \tag{B.12}$$

If we equate Equations B.5 and B.12, then we end up with a a relation between the spherical harmonic electric field coefficients and the sky electric field co-efficients:

$$E_{lm} = 2\pi(-i)^l e_{lm} \tag{B.13}$$

Next up is to calculate E_{lm}. At this point we use the method in Carozzi (2015) directly. We model each measurement as a delta function displaced from the origin of our co-ordinate system:

$$E(r,k,\theta,\phi) = \sum_{i=1}^{N} \frac{E_i(k_0)}{r^2 \sin(\theta)} \delta(r - r_i)\delta(k - k_0)\delta(\theta - \theta_i)\delta(\phi - \phi_i) \tag{B.14}$$

The integral is taken over the volume bounding the visibilities, after first multiplying by $j_l(k_0 r) Y^*_{lm}(\theta_i, \phi_i)$:

$$\int_0^{\infty}\int_0^{\pi}\int_0^{2\pi} \sum_{lm}\sum_{i=1}^{N} \frac{E_i(k_0)}{r^2 \sin(\theta)} \delta(r - r_i)\delta(k - k_0)\delta(\theta - \theta_i)\delta(\phi - \phi_i) r^2 \sin(\theta) dr\, d\theta\, d\phi \tag{B.15}$$

$E(k, r, \theta, \phi)$ naturally reduces as a $1/r$ law, but we neglect this term in this case as we only care about the angular distribution. The integral above becomes:

$$\sum_{i=1}^{N} \sum_{l,m} E_i(k_0) j_l(k_0 r) Y_{lm}^*(\theta_i, \phi_i) \delta(k - k_0) \tag{B.16}$$

Now we do the same treatment for the left hand sideof Equation B.14, using Equation B.5, the LHS becomes:

$$\int_0^\infty \int_0^\pi \int_0^{2\pi} \sum_{l,m} E_{lm} Y_{lm}(\Omega_k) j_l k r Y_{lm}^*(\Omega_k) j_l(k_0 r) dr \, d\theta \, d\phi \tag{B.17}$$

Which becomes:

$$\sum_{l,m} E_{lm} \delta_{ll} \delta_{mm} \int_0^\infty j_l(kr) j_l(k_0 r) dr \tag{B.18}$$

$$= \sum_{l,m} E_{lm} \frac{\pi \delta(k - k_0)}{2k_0^2} \tag{B.19}$$

Then we integrate over all k and equate the LHS and RHS to calculate:

$$E_{lm} = \frac{k_0^2}{\pi} \sum_{i=1}^{N} E_i(k_0) j_l(k_0 r_i) Y_{lm}^*(\theta_i, \phi_i) \tag{B.20}$$

We can then recover the sky co-efficients from the relationship in Equation B.13. The key result here is that a relationship exists between the sky and the electric field measurements at each antenna, in the same manner as in the EPIC correlator described in Thyagarajan et al. (2017); Kent et al. (2019a,b).

B.2 Practicality

A subtlety lies in that to perform statistical averaging to reduce thermal noise, as is commonly done with a visibility in a correlator, the electric field coefficients must be first projected onto the celestial sphere, then multiplied by the complex conjugate of the projection, as indicated by Equation B.3.

In a radio interferometer, the electric fields represent the fourier transform of the digitized voltages from each antenna. Thus to perform imaging of the sky using antenna electric fields using this method the spherical harmonic coefficients must be expanded in real-time, then projected onto the celestial sphere. This projection is then averaged over time to achieve the statistical reduction of thermal noise (see Appendix A).

The complexity of a spherical harmonics expansion of the visibilities is of order $\mathbb{O}(N_A^2 L^3)$, where N_A is the number of antennas and L is the spherical harmonic degree. For an electric field expansion this naturally reduces to $\mathbb{O}(N_A L^3)$, however with the difficulty that it is now necessary to do this expansion and projection in real-time in a radio interferometer. This is less feasible than other schemes of wide-field direct imaging, and therefore not practical with the current computational hardware available. It has been noted by McEwen and Scaife (2008) that it is additionally possible to solve the problem of simulation of visibilities by expansion of the sky brightness distribution into a set of Spherical Harr Wavelet (SHW) coefficients. Whilst this particular expansion was not considered for solving the wide field imaging problem for electric fields, it may be a more numerically tractable method. McEwen and Scaife (2008) noted that the SHW expansion is more well-posed, and has greater numerical stability as a result. They are easier to compute to higher orders, with greater benefits in terms of computational complexity, and may be a more robust estimator of the sky brightness distribution. In future work it would be advisable to see if the SHW expansion might be applicable to the problem of direct wide-field imaging. The requirement to be imaged in real-time remains however, and poses a significant challenge for any method to solve for this unique wide-field imaging problem.

OPTIMISED CONVOLUTION KERNELS

C.1 Reference C++ Kernel

```
#define ILP 16
static inline std::complex<double> deconvolve_visibility_(double u, double v, double w,
                                        double du, double dw,
                                        int aa_support_uv, int aa_support_w,
                                        int oversampling, int oversampling_w,
                                        int w_planes, int grid_size,
                                        const vector3D<std::complex<double> >& wstacks,
                                        struct sep_kernel_data *grid_conv_uv,
                                        struct sep_kernel_data *grid_conv_w){

    // Begin De-convolution process using Sze-Tan Kernels.
    std::complex<double> vis_sze = {0.0,0.0};

    // U/V/W oversample values
    double flu = u - std::ceil(u/du)*du;
    double flv = v - std::ceil(v/du)*du;
    double flw = w - std::ceil(w/dw)*dw;

    int ovu = static_cast<int>(std::floor(std::abs(flu)/du * oversampling));
    int ovv = static_cast<int>(std::floor(std::abs(flv)/du * oversampling));
    int ovw = static_cast<int>(std::floor(std::abs(flw)/dw * oversampling_w));

    int aa_h = std::floor(aa_support_uv/2);
    int aaw_h = std::floor(aa_support_w/2);

    for(int dwi = -aaw_h; dwi < aaw_h; ++dwi){

      int dws = static_cast<int>(std::ceil(w/dw)) +
      static_cast<int>(std::floor(w_planes/2)) + dwi;
        int aas_w = aa_support_w * ovw + (dwi+aaw_h);
        double gridconv_w = grid_conv_w->data[aas_w];

        for(int dvi = -aa_h; dvi < aa_h; ++dvi){

            int dvs = static_cast<int>(std::ceil(v/du)) + grid_size + dvi;
            int aas_v = aa_support_uv * ovv + (dvi+aa_h);
```

```
            double gridconv_uv = gridconv_w * grid_conv_uv->data[aas_v];

            for(int dui = -aa_h; dui < aa_h; ++dui){

                int dus = static_cast<int>(std::ceil(u/du)) + grid_size + dui;
                int aas_u = aa_support_uv * ovu + (dui+aa_h);
                double gridconv_u = grid_conv_uv->data[aas_u];
                double gridconv_uvw = gridconv_uv * gridconv_u;
                vis_sze += (wstacks(dus,dvs,dws) * gridconv_uvw );
            }
        }
    }

    return vis_sze;
}
```

C.2 General CUDA Kernel

```
#define ILP 16
template <typename FloatType>
__global__ void deconvolve_3D(thrust::complex<FloatType> *wstacks,thrust::complex<FloatType> *vis,
                              FloatType *uvec, FloatType *vvec, FloatType *wvec,
                              FloatType *gcf_uv, FloatType *gcf_w, FloatType du, FloatType dw,
                              int vis_num, int aa_support_uv, int aa_support_w,
                              int oversampling, int oversampling_w,
                              int w_planes, int grid_size, int oversampg){

    const int x = ILP*(blockIdx.x * blockDim.x + threadIdx.x);
    if(x+ILP >= vis_num) return;

    const int aa_h = aa_support_uv/2;
    const int aaw_h = aa_support_w/2;

    for(int i = 0; i < ILP; i++){

        if (i+x > vis_num) continue;
        FloatType u = uvec[i+x];
        FloatType v = vvec[i+x];
        FloatType w = wvec[i+x];

        FloatType flu = u - cuda_ceil(u/du)*du;
        FloatType flv = v - cuda_ceil(v/du)*du;
        FloatType flw = w - cuda_ceil(w/dw)*dw;

        int ovu = static_cast<int>(cuda_floor(cuda_fabs(flu)/du * oversampling));
        int ovv = static_cast<int>(cuda_floor(cuda_fabs(flv)/du * oversampling));
        int ovw = static_cast<int>(cuda_floor(cuda_fabs(flw)/dw * oversampling_w));

        thrust::complex<FloatType> vis_accum = {0.0,0.0};

        for(int dwi = -aaw_h; dwi < aaw_h; ++dwi){

            int dws = static_cast<int>(cuda_ceil(w/dw)) +
                w_planes/2 + dwi;
            int aas_w = aa_support_w * ovw + (dwi + aaw_h);
            FloatType gridconv_w = gcf_w[aas_w];


            for(int dui = -aa_h; dui < aa_h; ++dui){
                int dus = static_cast<int>(cuda_ceil(u/du)) + grid_size + dui;
                int aas_u = aa_support_uv * ovu + (dui + aa_h);
                FloatType gridconv_uw = gridconv_w * gcf_uv[aas_u];

                for(int dvi = -aa_h; dvi < aa_h; ++dvi){

                    int dvs = static_cast<int>(cuda_ceil(v/du)) + grid_size + dvi;
```

```
                    int aas_v = aa_support_uv * ovv + (dvi + aa_h);
                    FloatType gridconv_uvw = gridconv_uw * gcf_uv[aas_v];
                    thrust::complex<FloatType> conv_point = wstacks[dws * oversampg * oversampg
                    + dus * oversampg + dvs];
                    vis_accum += (conv_point * gridconv_uvw);
                }
            }
        }
        vis[i+x] = vis_accum;
    }
}
```

C.3 Optimised CUDA Kernel

```
#define ILP 16
template <typename FloatType>
__global__ void deconvolve_3D_wp_2(thrust::complex<FloatType>* __restrict__ wstacks,
                                   thrust::complex<FloatType>* vis,
                                   FloatType* __restrict__ uvec,
                                   FloatType* __restrict__ vvec,
                                   FloatType* __restrict__ wvec,
                                   FloatType* __restrict__ gcf_uv, FloatType* __restrict__ gcf_w,
                                   FloatType du, FloatType dw,
                                   int vis_num, int aa_support_uv, int aa_support_w,
                                   int oversampling, int oversampling_w,
                                   int w_planes, int grid_size, int oversampg){


    const int aa_h = aa_support_uv/2;
    const int aaw_h = aa_support_w/2;

    // Warp Information

    const unsigned int warp_total = blockDim.x / 32;
    const unsigned int warp_idx = threadIdx.x / 32;
    const unsigned int lane_idx = threadIdx.x 32;
    const unsigned int quarter_warp_idx = lane_idx / 8;
    const unsigned int quarter_warp_lane_idx = lane_idx 8;
    // Shared memory
    extern __shared__ unsigned int array[];

    // Offset for each warp in the u/v/w vectors
    const unsigned int uvw_offset = ILP*(blockIdx.x * warp_total + warp_idx);
    // Start pointer for the saved visibility values
    thrust::complex<FloatType> *vis_space =
        reinterpret_cast<thrust::complex<FloatType>*>(array) + warp_idx * 32;

    for(int i = 0; i < ILP; ++i){

        FloatType u = uvec[uvw_offset + i];
        FloatType v = vvec[uvw_offset + i];
        FloatType w = wvec[uvw_offset + i];

        FloatType flu = u - cuda_ceil(u/du)*du;
        FloatType flv = v - cuda_ceil(v/du)*du;
        FloatType flw = w - cuda_ceil(w/dw)*dw;

        int ovu = static_cast<int>(cuda_floor(cuda_fabs(flu)/du * oversampling));
        int ovv = static_cast<int>(cuda_floor(cuda_fabs(flv)/du * oversampling));
        int ovw = static_cast<int>(cuda_floor(cuda_fabs(flw)/dw * oversampling_w));

        int aas_u = aa_support_uv * ovu + quarter_warp_lane_idx;
        int aas_w = aa_support_w * ovw + quarter_warp_idx;

        int w_grid = static_cast<int>(cuda_ceil(w/dw)) + w_planes/2 - aaw_h
            + quarter_warp_idx;
```

```
        int u_grid = static_cast<int>(cuda_ceil(u/du)) + grid_size - aa_h + quarter_warp_lane_idx;
        int v_grid_p = static_cast<int>(cuda_ceil(v/du)) + grid_size - aa_h;
        int grid_coord = w_grid * oversampg * oversampg  + u_grid;
        int aas_v_p = aa_support_uv * ovv;


        /*
           COMPUTATION: Each thread in the warp accumulates a certain number of contributions to
           the grid point.

           Each thread operates within its quarter-warp. This is an optimisation based on coalescing
           memory accesses to the grid and kernels. Convolution size of 8*8*4 means we can work
           on lines of 8 elements at a time.

           After this per-thread accumulation, a shuffle down operation is performed across the
           warp.

        */


        thrust::complex<FloatType> value = {0.0,0.0};
        // CUDA Compiler automatically unrolls this at -O3
        FloatType conv_value_pre = 1.0 * gcf_w[aas_w] * gcf_uv[aas_u];
        for(int lb_v = 0; lb_v < 8; ++lb_v){
            thrust::complex<FloatType> grid = wstacks[grid_coord + (v_grid_p+lb_v) * oversampg];
            FloatType conv_value = conv_value_pre * gcf_uv[aas_v_p + lb_v];
            value += grid * conv_value;//
        }

        FloatType realn = value.real();
        FloatType imagn = value.imag();
        __syncwarp(); // Synchronise warp before reduction.

        // Warp shuffle reduction
        for(int offset = 16; offset > 0; offset /= 2){
            realn += __shfl_down_sync(0xFFFFFFFF,realn,offset);
            imagn += __shfl_down_sync(0xFFFFFFFF,imagn,offset);
        }
        if(lane_idx == 0){
            thrust::complex<FloatType> temp = {realn, imagn};
            vis_space[i] = temp;
        }
        __syncwarp();
    }
    __syncwarp();
    vis[uvw_offset + lane_idx] = vis_space[lane_idx];
}
```

Bibliography

Adami, K. Z. and Turner, W. (2011). SKA Station Beamformer Concept Description. Technical report, Science Data Processor Consortium.

Amiri, M., Bandura, K., Bhardwaj, M., Boubel, P., Boyce, M. M., Boyle, P. J., Brar, C., Burhanpurkar, M., Cassanelli, T., Chawla, P., Cliche, J. F., Cubranic, D., Deng, M., Denman, N., Dobbs, M., Fandino, M., Fonseca, E., Gaensler, B. M., Gilbert, A. J., Gill, A., Giri, U., Good, D. C., Halpern, M., Hanna, D. S., Hill, A. S., Hinshaw, G., Höfer, C., Josephy, A., Kaspi, V. M., Landecker, T. L., Lang, D. A., Lin, H. H., Masui, K. W., Mckinven, R., Mena-Parra, J., Merryfield, M., Michilli, D., Milutinovic, N., Moatti, C., Naidu, A., Newburgh, L. B., Ng, C., Patel, C., Pen, U. L., Pinsonneault-Marotte, T., Pleunis, Z., Rafiei-Ravandi, M., Rahman, M., Ransom, S. M., Renard, A., Scholz, P., Shaw, J. R., Siegel, S. R., Smith, K. M., Stairs, I. H., Tendulkar, S. P., Tretyakov, I., Vanderlinde, K., Yadav, P., and Collaboration, T. C. (2019a). A second source of repeating fast radio bursts. *Nature*.

Amiri, M., Bandura, K., Bhardwaj, M., Boubel, P., Boyce, M. M., Boyle, P. J., Brar, C., Burhanpurkar, M., Chawla, P., Cliche, J. F., Cubranic, D., Deng, M., Denman, N., Dobbs, M., Fandino, M., Fonseca, E., Gaensler, B. M., Gilbert, A. J., Giri, U., Good, D. C., Halpern, M., Hanna, D., Hill, A. S., Hinshaw, G., Höfer, C., Josephy, A., Kaspi, V. M., Landecker, T. L., Lang, D. A., Masui, K. W., Mckinven, R., Mena-Parra, J., Merryfield, M., Milutinovic, N., Moatti, C., Naidu, A., Newburgh, L. B., Ng, C., Patel, C., Pen, U., Pinsonneault-Marotte, T., Pleunis, Z., Rafiei-Ravandi, M., Ransom, S. M., Renard, A., Scholz, P., Shaw, J. R., Siegel, S. R., Smith, K. M., Stairs, I. H., Tendulkar, S. P., Tretyakov, I., Vanderlinde, K., Yadav, P., and The CHIME/FRB Collaboration (2019b). Observations of fast radio bursts at frequencies down to 400 megahertz. *Nature*, 566(7743):230–234.

Baldwin, J. E., Haniff, C. A., Mackay, C. D., and Warner, P. J. (1986). Closure phase in high-resolution optical imaging. *Nature*, 320(6063):595–597.

Bandura, K., Addison, G. E., Amiri, M., Bond, J. R., Campbell-Wilson, D., Connor, L., Cliche, J.-F., Davis, G., Deng, M., Denman, N., Dobbs, M., Fandino, M., Gibbs, K., Gilbert, A., Halpern, M., Hanna, D., Hincks, A. D., Hinshaw, G., Hofer, C., Klages, P., Landecker, T. L., Masui, K., Mena, J., Newburgh, L. B., Pen, U.-L., Peterson, J. B., Recnik, A., Shaw, J. R.,

Sigurdson, K., Sitwell, M., Smecher, G., Smegal, R., Vanderlinde, K., and Wiebe, D. (2014). Canadian Hydrogen Intensity Mapping Experiment (CHIME) Pathfinder. *Proceedings of the SPIE*, 9145.

Barott, W. C., Milgrome, O., Wright, M., MacMahon, D., Kilsdonk, T., Backus, P. R., and Dexter, M. (2011). Real-time beamforming using high-speed FPGAs at the Allen Telescope Array. *Radio Science*, 46.

Beardsley, A. P., Thyagarajan, N., Bowman, J. D., and Morales, M. F. (2017). An Efficient Feedback Calibration Algorithm for Direct Imaging Radio Telescopes. *Monthly Notices of the Royal Astronomical Society*, 470(4):4720–4731. arXiv: 1603.02126.

Bhatnagar, S., Cornwell, T. J., Golap, K., and Uson, J. M. (2008). Correcting direction-dependent gains in the deconvolution of radio interferometric images. *Astronomy & Astrophysics*, 487(1):419–429. arXiv: 0805.0834.

Blackburn, L., Pesce, D. W., Johnson, M. D., Wielgus, M., Chael, A. A., Christian, P., and Doeleman, S. S. (2019). Closure statistics in radio interferometric data. *arXiv:1910.02062 [astro-ph]*. arXiv: 1910.02062.

Bolton, R., Broekema, P., Cornwell, T., van Diepen, G., Hollitt, C., Johnston-Hollitt, M., Levin Preston, L., Mika, A., Nijboer, R., Nikolic, B., Salvini, S., Rampadarath, H., Scaife, A., Stappers, B., and Wortmann, P. (2019). Parametric models of SDP compute requirements.

Born, M. and Wolf, E. (1999). *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Cambridge University Press, Cambridge, 7th (expanded) ed edition.

Bunton, J. D. (2011). Antenna Array Geometries to Reduce the Compute Load in Radio Telescopes. *IEEE Transactions on Antennas and Propagation*, 59(6):2041–2046.

Caleb, M., Flynn, C., Bailes, M., Barr, E. D., Bateman, T., Bhandari, S., Campbell-Wilson, D., Farah, W., Green, A. J., Hunstead, R. W., Jameson, A., Jankowski, F., Keane, E. F., Parthasarathy, A., Ravi, V., Rosado, P. A., van Straten, W., and Krishnan, V. V. (2017). The first interferometric detections of Fast Radio Bursts. *Monthly Notices of the Royal Astronomical Society*, 468(3):3746–3756. arXiv: 1703.10173.

Carilli, C., Thyagarajan, N., Kent, J., Nikolic, B., Gale-Sides, K., Kern, N. S., Bernardi, G., Mesinger, A., Matika, S., and Collaboration, T. H. (2020). Imaging and Modeling Data from the Hydrogen Epoch of Reionization Array. *arXiv:2002.07692*.

Carilli, C. L., Nikolic, B., Thyagarajan, N., Gale-Sides, K., Abdurashidova, Z., Aguirre, J. E., Alexander, P., Ali, Z. S., Balfour, Y., Beardsley, A. P., Bernardi, G., Bowman, J. D., Bradley, R. F., Burba, J., Cheng, C., DeBoer, D. R., Dexter, M., de~Lera~Acedo, E., Dillon, J. S., Ewall-Wice, A., Fadana, G., Fagnoni, N., Fritz, R., Furlanetto, S. R., Ghosh, A., Glendenning, B., Greig, B., Grobbelaar, J., Halday, Z., Hazelton, B. J., Hewitt, J. N., Hickish, J., Jacobs,

D. C., Julius, A., Kariseb, M., Kohn, S. A., Kolopanis, M., Lekalake, T., Liu, A., Loots, A., MacMahon, D., Malan, L., Malgas, C., Maree, M., Martinot, Z., Matsetela, E., Mesinger, A., Molewa, M., Morales, M. F., Neben, A. R., Parsons, A. R., Patra, N., Pieterse, S., La Plante, P., Pober, J. C., Razavi-Ghods, N., Ringuette, J., Robnett, J., Rosie, K., Sell, R., Sims, P., Smith, C., Syce, A., Williams, P. K. ~., and Zheng, H. (2018). HI 21cm Cosmology and the Bi-spectrum: Closure Diagnostics in Massively Redundant Interferometric Arrays. *Radio Science*, 53(6):845–865. arXiv: 1805.00953.

Carozzi, T. D. (2015). Imaging on a Sphere with Interferometers: the Spherical Wave Harmonic Transform. *Monthly Notices of the Royal Astronomical Society: Letters*, 451(1):L6–L10. arXiv: 1504.04485.

Chime/Frb Collaboration, Amiri, M., Bandura, K., Berger, P., Bhardwaj, M., Boyce, M. M., Boyle, P. J., Brar, C., Burhanpurkar, M., Chawla, P., Chowdhury, J., Cliche, J.-F., Cranmer, M. D., Cubranic, D., Deng, M., Denman, N., Dobbs, M., Fandino, M., Fonseca, E., Gaensler, B. M., Giri, U., Gilbert, A. J., Good, D. C., Guliani, S., Halpern, M., Hinshaw, G., Höfer, C., Josephy, A., Kaspi, V. M., Landecker, T. L., Lang, D., Liao, H., Masui, K. W., Mena-Parra, J., Naidu, A., Newburgh, L. B., Ng, C., Patel, C., Pen, U.-L., Pinsonneault-Marotte, T., Pleunis, Z., Rafiei Ravandi, M., Ransom, S. M., Renard, A., Scholz, P., Sigurdson, K., Siegel, S. R., Smith, K. M., Stairs, I. H., Tendulkar, S. P., Vanderlinde, K., and Wiebe, D. V. (2018). The CHIME Fast Radio Burst Project: System Overview. *The Astrophysical Journal*, 863(1):48.

Cooley, J. W. and Tukey, J. W. (1965). An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation*, 19(90):297–301.

Cornwell, T. J., Golap, K., and Bhatnagar, S. (2008). The noncoplanar baselines effect in radio interferometry: The W-projection algorithm. *IEEE Journal of Selected Topics in Signal Processing*, 2(5):647–657.

Cornwell, T. J. and Perley, R. A. (1992). Radio-interferometric imaging of very large fields-The problem of non-coplanar arrays. *Astronomy and Astrophysics*, 261:353–364.

Cranmer, M. D., Barsdell, B. R., Price, D. C., Dowell, J., Garsden, H., Dike, V., Eftekhari, T., Hegedus, A. M., Malins, J., Obenberger, K. S., Schinzel, F., Stovall, K., Taylor, G. B., and Greenhill, L. J. (2017). Bifrost: a Python/C++ Framework for High-Throughput Stream Processing in Astronomy. *Journal of Astronomical Instrumentation*, 6(04):1750007. arXiv: 1708.00720.

Daishido, T., Asuma, K., Nishibori, K., Nakajima, J., Yano, M., Otobe, E., Watanabe, N., Tsuchiya, A., and Iwase, S. (1991). Direct imaging digital lens for transient radio source survey. In Cornwell, T. J. and Perley, R. A., editors, *IAU Colloq. 131: Radio Interferometry. Theory, Techniques, and Applications*, volume 19 of *Astronomical Society of the Pacific Conference Series*, pages 86–89. Cambridge University Press.

DeBoer, D. R., Parsons, A. R., Aguirre, J. E., Alexander, P., Ali, Z. S., Beardsley, A. P., Bernardi, G., Bowman, J. D., Bradley, R. F., Carilli, C. L., Cheng, C., Acedo, E. d. L., Dillon, J. S., Ewall-Wice, A., Fadana, G., Fagnoni, N., Fritz, R., Furlanetto, S. R., Glendenning, B., Greig, B., Grobbelaar, J., Hazelton, B. J., Hewitt, J. N., Hickish, J., Jacobs, D. C., Julius, A., Kariseb, M., Kohn, S. A., Lekalake, T., Liu, A., Loots, A., MacMahon, D., Malan, L., Malgas, C., Maree, M., Mathison, N., Matsetela, E., Mesinger, A., Morales, M. F., Neben, A. R., Patra, N., Pieterse, S., Pober, J. C., Razavi-Ghods, N., Ringuette, J., Robnett, J., Rosie, K., Sell, R., Smith, C., Syce, A., Tegmark, M., Thyagarajan, N., Williams, P. K. G., and Zheng, H. (2017). Hydrogen Epoch of Reionization Array (HERA). *Publications of the Astronomical Society of the Pacific*, 129(974):045001. arXiv: 1606.07473.

Dewdney, P., Hall, P., Schilizzi, R., and Lazio, T. (2009a). The Square Kilometre Array. *Proceedings of the IEEE*, 97(8):1482–1496.

Dewdney, P. E., Hall, P. J., Schilizzi, R. T., and Lazio, T. J. L. W. (2009b). The Square Kilometre Array. *Proceedings of the IEEE*, 97(8):1482–1496.

Dillon, J. (2018). IDR2.1 Memo. Technical report, HERA Collaboration.

Dowell, J. and Taylor, G. B. (2018). The Swarm Telescope Concept. *Journal of Astronomical Instrumentation*, 07(02n03):1850006. arXiv: 1806.10634.

Dowell, J., Wood, D., Stovall, K., Ray, P. S., Clarke, T., and Taylor, G. (2012). The long wavelength array software library. *Journal of Astronomical Instrumentation*, 01(01):1250006.

Ellingson, S. W., Taylor, G. B., Craig, J., Hartman, J., Dowell, J., Wolfe, C. N., Clarke, T. E., Hicks, B. C., Kassim, N. E., Ray, P. S., Rickard, L. J., Schinzel, F. K., and Weiler, K. W. (2013). The LWA1 Radio Telescope. *IEEE Transactions on Antennas and Propagation*, 61:2540–2549.

Fagnoni, N., Acedo, E. d. L., DeBoer, D. R., Abdurashidova, Z., Aguirre, J. E., Alexander, P., Ali, Z. S., Balfour, Y., Beardsley, A. P., Bernardi, G., Billings, T. S., Bowman, J. D., Bradley, R. F., Bull, P., Burba, J., Carilli, C. L., Cheng, C., Dexter, M., Dillon, J. S., Ewall-Wice, A., Fritz, R., Furlanetto, S. R., Gale-Sides, K., Glendenning, B., Gorthi, D., Greig, B., Grobbelaar, J., Halday, Z., Hazelton, B. J., Hewitt, J. N., Hickish, J., Jacobs, D. C., Josaitis, A., Julius, A., Kern, N. S., Kerrigan, J., Kim, H., Kittiwisit, P., Kohn, S. A., Kolopanis, M., Lanman, A., La Plante, P., Lekalake, T., Liu, A., MacMahon, D., Malan, L., Malgas, C., Maree, M., Martinot, Z. E., Matsetela, E., Parra, J. M., Mesinger, A., Molewa, M., Morales, M. F., Mosiane, T., Neben, A. R., Nikolic, B., Parsons, A. R., Patra, N., Pieterse, S., Pober, J. C., Razavi-Ghods, N., Robnett, J., Rosie, K., Sims, P., Smith, C., Syce, A., Thyagarajan, N., Williams, P. K. G., and Zheng, H. (2019). Electrical and electromagnetic co-simulations of the HERA Phase I receiver system including the effects of mutual coupling, and impact

on the EoR window. *arXiv:1908.02383 [astro-ph]*. arXiv: 1908.02383.

Feller, W. (1945). The fundamental limit theorems in probability. *Bulletin of the American Mathematical Society*, 51(11):800–832.

Feynman, R. P., Leighton, R. B., and Sands, M. L. (1965). *Feynman lectures on physics. Volume 3: Quantum mechanics*, volume 3. California Institute of Technology.

Foster, G., Hickish, J., Magro, A., Price, D. C., and Adami, K. Z. (2014). Implementation of a Direct-Imaging and FX Correlator for the BEST-2 Array. *Monthly Notices of the Royal Astronomical Society*, 439(3):3180–3188. arXiv: 1401.6753.

Frigo, M. and Johnson, S. (2005). The Design and Implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231.

Furlanetto, S. R. (2016). The 21-cm Line as a Probe of Reionization. In *Understanding the Epoch of Cosmic Reionization*, pages 247–280. Springer.

Green, D. A. (2011). A colour scheme for the display of astronomical intensity images. *Bulletin of the Astronomical Society of India*, 39:289–295.

Gull, S. F. and Skilling, J. (1984). Maximum entropy method in image processing. *IEE Proceedings F - Communications, Radar and Signal Processing*, 131(6):646–659.

Hazelton, B. J., Jacobs, D. C., Pober, J. C., and Beardsley, A. P. (2017). pyuvdata: an interface for astronomical interferometeric datasets in python. *The Journal of Open Source Software*, 2:140.

Henning, P., Ellingson, S. W., Taylor, G. B., Craig, J., Pihlstrom, Y., Rickard, L. J., Clarke, T. E., Kassim, N. E., and Cohen, A. (2010). The First Station of the Long Wavelength Array. *Proceedings, 3rd International SKA Forum, Science Meeting on Square kilometre array (ISKAF 2010): Assen, Netherlands, June 10-14, 2010*, ISKAF2010:024.

Högbom, J. A. (1974). Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines. *Astronomy and Astrophysics Supplement*, 15:417.

Jackson, J. D. (2007). *Classical electrodynamics*. John Wiley & Sons.

Jackson, J. I., Meyer, C. H., Nishimura, D. G., and Macovski, A. (1991). Selection of a convolution function for Fourier inversion using gridding (computerised tomography application). *IEEE Transactions on Medical Imaging*, 10(3):473–478.

Jennison, R. C. (1958). A phase sensitive interferometer technique for the measurement of the Fourier transforms of spatial brightness distributions of small angular extent. *Monthly Notices of the Royal Astronomical Society*, 118:276.

Kent, J., Beardsley, A. P., Bester, L., Gull, S. F., Nikolic, B., Dowell, J., Thyagarajan, N., Taylor, G. B., and Bowman, J. (2019a). Direct wide-field radio imaging in real-time at high time resolution using antenna electric fields. *Monthly Notices of the Royal Astronomical Society*, 491(1):254–263.

Kent, J., Dowell, J., Beardsley, A., Thyagarajan, N., Taylor, G., and Bowman, J. (2019b). A real-time, all-sky, high time resolution, direct imager for the long wavelength array. *Monthly Notices of the Royal Astronomical Society*, 486(4):5052–5060.

Kent, J. and Nikolic, B. (2016). Science Data Processor Memo 19: Quantify Power Efficiency of FFTs on NVIDIA GPUs.

Kortenkamp, S. J. and Dermott, S. F. (1998). Accretion of Interplanetary Dust Particles by the Earth. *ıcarus*, 135:469–495.

Kulkarni, S. R. (1989). Self-noise in interferometers - Radio and infrared. *The Astronomical Journal*, 98:1112.

Lam, M. D., Rothberg, E. E., and Wolf, M. E. (1991). The cache performance and optimizations of blocked algorithms. In *ACM SIGARCH Computer Architecture News*, volume 19, pages 63–74. ACM.

Landau, H. J. and Pollak, H. O. (1961). Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - II. *Bell System Technical Journal*, 40(1):65–84.

Li, W., Pober, J. C., Hazelton, B. J., Barry, N., Morales, M. F., Sullivan, I., Parsons, A. R., Ali, Z. S., Dillon, J. S., Beardsley, A. P., Bowman, J. D., Briggs, F., Byrne, R., Carroll, P., Crosse, B., Emrich, D., Ewall-Wice, A., Feng, L., Franzen, T. M. O., Hewitt, J. N., Horsley, L., Jacobs, D. C., Johnston-Hollitt, M., Jordan, C., Joseph, R. C., Kaplan, D. L., Kenney, D., Kim, H., Kittiwisit, P., Lanman, A., Line, J., McKinley, B., Mitchell, D. A., Murray, S., Neben, A., Offringa, A. R., Pallot, D., Paul, S., Pindor, B., Procopio, P., Rahimi, M., Riding, J., Sethi, S. K., Shankar, N. U., Steele, K., Subrahmanian, R., Tegmark, M., Thyagarajan, N., Tingay, S. J., Trott, C., Walker, M., Wayth, R. B., Webster, R. L., Williams, A., Wu, C., and Wyithe, S. (2018). Comparing Redundant and Sky-model-based Interferometric Calibration: A First Look with Phase II of the MWA. *The Astrophysical Journal*, 863(2):170.

Mardia, K. V. and Jupp, P. (2008). Basic Concepts and Models. In *Directional Statistics*, pages 25–56. Wiley-Blackwell.

Masui, K. W., Shaw, J. R., Ng, C., Smith, K. M., Vanderlinde, K., and Paradise, A. (2019). Algorithms for FFT Beamforming Radio Interferometers. *The Astrophysical Journal*, 879(1):16.

McEwen, J. D. and Scaife, A. M. M. (2008). Simulating full-sky interferometric observations. *Monthly Notices of the Royal Astronomical Society*, 389(3):1163–1178.

Mesinger, A. (2017). Reionization and Cosmic Dawn: theory and simulations. *Proceedings of the International Astronomical Union*, 12(S333):3–11. arXiv: 1801.02649.

Mesinger, A., Furlanetto, S., and Cen, R. (2011). 21cmFAST: A Fast, Semi-Numerical Simulation of the High-Redshift 21-cm Signal. *Monthly Notices of the Royal Astronomical Society*, 411(2):955–972. arXiv: 1003.3878.

Michelson, A. A. and Pease, F. G. (1921). Measurement of the Diameter of alpha-Orionis with the Interferometer. *The Astrophysical Journal*, 53:249–259.

Mitchell, D. and Bernardi, G. (2014). Science Data Processor Memo: Analysis of w-projection kernel size.

Mol, J. D. and Romein, J. W. (2011). The LOFAR Beam Former: Implementation and Performance Analysis. In Jeannot, E., Namyst, R., and Roman, J., editors, *Euro-Par 2011 Parallel Processing*, volume 6853, pages 328–339. Springer Berlin Heidelberg, Berlin, Heidelberg.

Morales, M. F. (2011). Enabling Next-Generation Dark Energy and Epoch of Reionization Radio Observatories with the MOFF Correlator. *Publications of the Astronomical Society of the Pacific*, 123(909):1265.

Morales, M. F. and Matejek, M. (2009). Software holography: interferometric data analysis for the challenges of next generation observatories. *Monthly Notices of the Royal Astronomical Society*, 400(4):1814–1820.

Mort, B., Dulwich, F., Salvini, S., Adami, K., and Jones, M. (2010). OSKAR: Simulating digital beamforming for the SKA aperture array. *IEEE International Symposium on Phased Array Systems and Technology*, pages 690–694.

Newburgh, L. B., Bandura, K., Bucher, M. A., Chang, T.-C., Chiang, H. C., Cliche, J. F., Davé, R., Dobbs, M., Clarkson, C., Ganga, K. M., Gogo, T., Gumba, A., Gupta, N., Hilton, M., Johnstone, B., Karastergiou, A., Kunz, M., Lokhorst, D., Maartens, R., Macpherson, S., Mdlalose, M., Moodley, K., Ngwenya, L., Parra, J. M., Peterson, J., Recnik, O., Saliwanchik, B., Santos, M. G., Sievers, J. L., Smirnov, O., Stronkhorst, P., Taylor, R., Vanderlinde, K., Vuuren, G. V., Weltman, A., and Witzemann, A. (2016). HIRAX: a probe of dark energy and radio transients. In *Ground-based and Airborne Telescopes VI*, volume 9906, pages 2039–2049. International Society for Optics and Photonics.

Nikolic, B., Small, D., and Kettenis, M. (2018). Minimal re-computation for exploratory data analysis in astronomy. *Astronomy and Computing*, 25:133 – 138.

Obenberger, K. S., Dowell, J. D., Malins, J., Parris, R. T., Pedersen, T. R., and Taylor, G. B. (2018). Using Lightning as a HF Signal Source to Produce Ionograms. *Radio Science*, 53(11):1419–1425.

Obenberger, K. S., Taylor, G. B., Hartman, J. M., Dowell, J., Ellingson, S. W., Helmboldt, J. F., Henning, P. A., Kavic, M., Schinzel, F. K., Simonetti, J. H., Stovall, K., and Wilson, T. L. (2014). Detection of Radio Emission from Fireballs. *The Astrophysical Journal*, 788(2):L26. arXiv: 1405.6772.

Offringa, A. R., McKinley, B., Hurley-Walker, N., Briggs, F. H., Wayth, R. B., Kaplan, D. L., Bell, M. E., Feng, L., Neben, A. R., Hughes, J. D., Rhee, J., Murphy, T., Bhat, N. D. R., Bernardi, G., Bowman, J. D., Cappallo, R. J., Corey, B. E., Deshpande, A. A.,

Emrich, D., Ewall-Wice, A., Gaensler, B. M., Goeke, R., Greenhill, L. J., Hazelton, B. J., Hindson, L., Johnston-Hollitt, M., Jacobs, D. C., Kasper, J. C., Kratzenberg, E., Lenc, E., Lonsdale, C. J., Lynch, M. J., McWhirter, S. R., Mitchell, D. A., Morales, M. F., Morgan, E., Kudryavtseva, N., Oberoi, D., Ord, S. M., Pindor, B., Procopio, P., Prabu, T., Riding, J., Roshi, D. A., Shankar, N. U., Srivani, K. S., Subrahmanyan, R., Tingay, S. J., Waterson, M., Webster, R. L., Whitney, A. R., Williams, A., and Williams, C. L. (2014). WSClean: an implementation of a fast, generic wide-field imager for radio astronomy. *Monthly Notices of the Royal Astronomical Society*, 444(1):606–619. arXiv: 1407.1943.

Ord, S. M., Tremblay, S. E., McSweeney, S. J., Bhat, N. D. R., Sobey, C., Mitchell, D. A., Hancock, P. J., and Kirsten, F. (2019). MWA tied-array processing I: Calibration and beamformation. *Publications of the Astronomical Society of Australia*, 36.

Parsons, A. R., Pober, J. C., Aguirre, J. E., Carilli, C. L., Jacobs, D. C., and Moore, D. F. (2012). A Per-baseline, Delay-spectrum Technique for Accessing the 21 cm Cosmic Reionization Signature. *The Astrophysical Journal*, 756(2):165.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., Bartolo, N., Battaner, E., Battye, R., Benabed, K., Benoit, A., Benoit-Levy, A., Bernard, J.-P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bonaldi, A., Bonavera, L., Bond, J. R., Borrill, J., Bouchet, F. R., Boulanger, F., Bucher, M., Burigana, C., Butler, R. C., Calabrese, E., Cardoso, J.-F., Catalano, A., Challinor, A., Chamballu, A., Chary, R.-R., Chiang, H. C., Chluba, J., Christensen, P. R., Church, S., Clements, D. L., Colombi, S., Colombo, L. P. L., Combet, C., Coulais, A., Crill, B. P., Curto, A., Cuttaia, F., Danese, L., Davies, R. D., Davis, R. J., de Bernardis, P., de Rosa, A., de Zotti, G., Delabrouille, J., Desert, F.-X., Di Valentino, E., Dickinson, C., Diego, J. M., Dolag, K., Dole, H., Donzelli, S., Dore, O., Douspis, M., Ducout, A., Dunkley, J., Dupac, X., Efstathiou, G., Elsner, F., Ensslin, T. A., Eriksen, H. K., Farhang, M., Fergusson, J., Finelli, F., Forni, O., Frailis, M., Fraisse, A. A., Franceschi, E., Frejsel, A., Galeotta, S., Galli, S., Ganga, K., Gauthier, C., Gerbino, M., Ghosh, T., Giard, M., Giraud-Heraud, Y., Giusarma, E., Gjerlow, E., Gonzalez-Nuevo, J., Gorski, K. M., Gratton, S., Gregorio, A., Gruppuso, A., Gudmundsson, J. E., Hamann, J., Hansen, F. K., Hanson, D., Harrison, D. L., Helou, G., Henrot-Versille, S., Hernandez-Monteagudo, C., Herranz, D., Hildebrandt, S. R., Hivon, E., Hobson, M., Holmes, W. A., Hornstrup, A., Hovest, W., Huang, Z., Huffenberger, K. M., Hurier, G., Jaffe, A. H., Jaffe, T. R., Jones, W. C., Juvela, M., Keihanen, E., Keskitalo, R., Kisner, T. S., Kneissl, R., Knoche, J., Knox, L., Kunz, M., Kurki-Suonio, H., Lagache, G., Lahteenmaki, A., Lamarre, J.-M., Lasenby, A., Lattanzi, M., Lawrence, C. R., Leahy, J. P., Leonardi, R., Lesgourgues, J., Levrier, F., Lewis, A., Liguori, M., Lilje, P. B., Linden-Vornle, M., Lopez-Caniego, M., Lubin, P. M., Macias-Perez, J. F.,

Maggio, G., Maino, D., Mandolesi, N., Mangilli, A., Marchini, A., Martin, P. G., Martinelli, M., Martinez-Gonzalez, E., Masi, S., Matarrese, S., Mazzotta, P., McGehee, P., Meinhold, P. R., Melchiorri, A., Melin, J.-B., Mendes, L., Mennella, A., Migliaccio, M., Millea, M., Mitra, S., Miville-Deschenes, M.-A., Moneti, A., Montier, L., Morgante, G., Mortlock, D., Moss, A., Munshi, D., Murphy, J. A., Naselsky, P., Nati, F., Natoli, P., Netterfield, C. B., Norgaard-Nielsen, H. U., Noviello, F., Novikov, D., Novikov, I., Oxborrow, C. A., Paci, F., Pagano, L., Pajot, F., Paladini, R., Paoletti, D., Partridge, B., Pasian, F., Patanchon, G., Pearson, T. J., Perdereau, O., Perotto, L., Perrotta, F., Pettorino, V., Piacentini, F., Piat, M., Pierpaoli, E., Pietrobon, D., Plaszczynski, S., Pointecouteau, E., Polenta, G., Popa, L., Pratt, G. W., Prezeau, G., Prunet, S., Puget, J.-L., Rachen, J. P., Reach, W. T., Rebolo, R., Reinecke, M., Remazeilles, M., Renault, C., Renzi, A., Ristorcelli, I., Rocha, G., Rosset, C., Rossetti, M., Roudier, G., d'Orfeuil, B. R., Rowan-Robinson, M., Rubino-Martin, J. A., Rusholme, B., Said, N., Salvatelli, V., Salvati, L., Sandri, M., Santos, D., Savelainen, M., Savini, G., Scott, D., Seiffert, M. D., Serra, P., Shellard, E. P. S., Spencer, L. D., Spinelli, M., Stolyarov, V., Stompor, R., Sudiwala, R., Sunyaev, R., Sutton, D., Suur-Uski, A.-S., Sygnet, J.-F., Tauber, J. A., Terenzi, L., Toffolatti, L., Tomasi, M., Tristram, M., Trombetti, T., Tucci, M., Tuovinen, J., Turler, M., Umana, G., Valenziano, L., Valiviita, J., Van Tent, B., Vielva, P., Villa, F., Wade, L. A., Wandelt, B. D., Wehus, I. K., White, M., White, S. D. M., Wilkinson, A., Yvon, D., Zacchei, A., and Zonca, A. (2016). Planck 2015 results. XIII. Cosmological parameters. *Astronomy & Astrophysics*, 594:A13. arXiv: 1502.01589.

Pratley, L., McEwen, J. D., d'Avezac, M., Cai, X., Perez-Suarez, D., Christidi, I., and Guichard, R. (2019). Distributed and parallel sparse convex optimization for radio interferometry with PURIFY. *arXiv:1903.04502 [astro-ph]*. arXiv: 1903.04502.

Prentice, J. P. M., Lovell, A. C. B., and Banwell, C. J. (1947). Radio echo observations of meteors. *Monthly Notices of the Royal Astronomical Society*, 107(2):155–163.

Rogers, A. E. E., Hinteregger, H. F., Whitney, A. R., Counselman, C. C., Shapiro, I. I., Wittels, J. J., Klemperer, W. K., Warnock, W. W., Clark, T. A., Hutton, L. K., Marandino, G. E., Ronnang, B. O., Rydbeck, O. E. H., and Niell, A. E. (1974). The structure of radio sources 3C 273B and 3C 84 deduced from the "closure" phases and visibility amplitudes observed with three-element interferometers. *The Astrophysical Journal*, 193:293–301.

Romein, J. W. (2012). An efficient work-distribution strategy for gridding radio-telescope data on GPUs. In *Proceedings of the 26th ACM international conference on Supercomputing*, pages 321–330. ACM.

Romein, J. W. (2016). A Comparison of Accelerator Architectures for Radio-Astronomical Signal-Processing Algorithms. In *2016 45th International Conference on Parallel Processing (ICPP)*, pages 484–489.

Romney, J. D. (1985). Lecture 4. Cross Correlators. In *Synthesis Imaging: Course Notes from an NRAO Summer School Held in Socorro, New Mexico.*, volume 6. Astronomical Society of the Pacific, San Francisco, CA.

Slepian, D. and Pollak, H. O. (1961). Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - I. *Bell System Technical Journal*, 40(1):43–63.

Tan, S. M. (1986). *Aperture Synthesis Mapping and Parameter Estimation*. PhD thesis, University of Cambridge, Cambridge.

Taylor, G., Dowell, J., Malins, J., Clarke, T., Kassim, N., Giacintucci, S., Hicks, B., Kooi, J., Peters, W., Polisensky, E., Schinzel, F., and Stovall, K. (2017). A Next Generation Low Band Observatory: A Community Study Exploring Low Frequency Options for ngVLA. *ArXiv e-prints*. arXiv:1708.00090.

Taylor, G. B., Carilli, C. L., and Perley, R. A., editors (1999). *Synthesis Imaging in Radio Astronomy II*, volume 180 of *Astronomical Society of the Pacific Conference Series*. IOP Publishing.

Taylor, G. B., Ellingson, S. W., Kassim, N. E., Craig, J., Dowell, J., Wolfe, C. N., Hartman, J., Bernardi, G., Clarke, T., Cohen, A., Dalal, N. P., Erickson, W. C., Hicks, B., Greenhill, L. J., Jacoby, B., Lane, W., Lazio, J., Mitchell, D., Navarro, R., Ord, S. M., Pihlstrom, Y., Polisensky, E., Ray, P. S., Rickard, L. J., Schinzel, F. K., Schmitt, H., Sigman, E., Soriano, M., Stewart, K. P., Stovall, K., Tremblay, S., Wang, D., Weiler, K. W., White, S., and Wood, D. L. (2012). First Light for the First Station of the Long Wavelength Array. *Journal of Astronomical Instrumentation*, 01(01):1250004. arXiv: 1206.6733.

Thompson, A. R., Moran, J. M., and Swenson, G. W. (2017). *Interferometry and Synthesis in Radio Astronomy*. Astronomy and Astrophysics Library. Springer International Publishing, Cham.

Thyagarajan, N., Beardsley, A. P., Bowman, J. D., and Morales, M. F. (2017). A generic and efficient E-field Parallel Imaging Correlator for next-generation radio telescopes. *Monthly Notices of the Royal Astronomical Society*, 467(1):715–730.

Thyagarajan, N., Carilli, C. L., and Nikolic, B. (2018). Detecting Cosmic Reionization Using the Bispectrum Phase. *Physical Review Letters*, 120(25).

Thyagarajan, N., Carilli, C. L., Nikolic, B., Kent, J., Mesinger, A., Kern, N. S., Bernardi, G., Matika, S., Abdurashidova, Z., Aguirre, J. E., Alexander, P., Ali, Z. S., Balfour, Y., Beardsley, A. P., Billings, T. S., Bowman, J. D., Bradley, R. F., Burba, J., Carey, S., Cheng, C., DeBoer, D. R., Dexter, M., Acedo, E. d. L., Dillon, J. S., Ely, J., Ewall-Wice, A., Fagnoni, N., Fritz, R., Furlanetto, S. R., Gale-Sides, K., Glendenning, B., Gorthi, D., Greig, B., Grobbelaar, J., Halday, Z., Hazelton, B. J., Hewitt, J. N., Hickish, J., Jacobs, D. C., Julius, A., Kerrigan, J., Kittiwisit, P., Kohn, S. A., Kolopanis, M., Lanman, A., La Plante, P., Lekalake, T., Lewis, D.,

Liu, A., MacMahon, D., Malan, L., Malgas, C., Maree, M., Martinot, Z. E., Matsetela, E., Molewa, M., Morales, M. F., Mosiane, T., Neben, A. R., Parsons, A. R., Patra, N., Pieterse, S., Pober, J. C., Razavi-Ghods, N., Ringuette, J., Robnett, J., Rosie, K., Sims, P., Smith, C., Syce, A., Williams, P. K. G., and Zheng, H. (2020). Detection of Cosmic Structures using the Bispectrum Phase. II. First Results from Application to Cosmic Reionization Using the Hydrogen Epoch of Reionization Array. *Physical Review D, Accepted*. arXiv: 2005.10275.

Tingay, S. J., Goeke, R., Bowman, J. D., Emrich, D., Ord, S. M., Mitchell, D. A., Morales, M. F., Booler, T., Crosse, B., Wayth, R. B., Lonsdale, C. J., Tremblay, S., Pallot, D., Colegate, T., Wicenec, A., Kudryavtseva, N., Arcus, W., Barnes, D., Bernardi, G., Briggs, F., Burns, S., Bunton, J. D., Cappallo, R. J., Corey, B. E., Deshpande, A., Desouza, L., Gaensler, B. M., Greenhill, L. J., Hall, P. J., Hazelton, B. J., Herne, D., Hewitt, J. N., Johnston-Hollitt, M., Kaplan, D. L., Kasper, J. C., Kincaid, B. B., Koenig, R., Kratzenberg, E., Lynch, M. J., Mckinley, B., Mcwhirter, S. R., Morgan, E., Oberoi, D., Pathikulangara, J., Prabu, T., Remillard, R. A., Rogers, A. E. E., Roshi, A., Salah, J. E., Sault, R. J., Udaya-Shankar, N., Schlagenhaufer, F., Srivani, K. S., Stevens, J., Subrahmanyan, R., Waterson, M., Webster, R. L., Whitney, A. R., Williams, A., Williams, C. L., and Wyithe, J. S. B. (2013). The Murchison Widefield Array: The Square Kilometre Array Precursor at Low Radio Frequencies. *Publications of the Astronomical Society of Australia*, 30.

Tol, S. v. d., Veenboer, B., and Offringa, A. R. (2018). Image Domain Gridding: a fast method for convolutional resampling of visibilities. *Astronomy & Astrophysics*, 616:A27.

Veenboer, B., Petschow, M., and Romein, J. W. (2017). Image-Domain Gridding on Graphics Processors. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 545–554.

Wayth, R. B., Tingay, S. J., Trott, C. M., Emrich, D., Johnston-Hollitt, M., McKinley, B., Gaensler, B. M., Beardsley, A. P., Booler, T., Crosse, B., Franzen, T. M. O., Horsley, L., Kaplan, D. L., Kenney, D., Morales, M. F., Pallot, D., Sleap, G., Steele, K., Walker, M., Williams, A., Wu, C., Cairns, I. H., Filipovic, M. D., Johnston, S., Murphy, T., Quinn, P., Staveley-Smith, L., Webster, R., and Wyithe, J. S. B. (2018). The Phase II Murchison Widefield Array: Design Overview. *Publications of the Astronomical Society of Australia*, 35:e033. arXiv: 1809.06466.

Wijnholds, S. J., Salvini, S., Dodson, R., Mort, B., Dulwich, F., Willis, A., and Still, J. (2016). Science Data Processor Memo 4: Feasibility analysis of baseline-dependent averaging.

Williams, S., Waterman, A., and Patterson, D. (2009). Roofline: An Insightful Visual Performance Model for Multicore Architectures. *Commun. ACM*, 52(4):65–76.

Ye, H. (2019). *Accurate image reconstruction in radio interferometry*. PhD thesis, University of Cambridge.

Ye, H., Gull, S. F., Tan, S. M., and Nikolic, B. (2019). Optimal gridding and degridding in radio interferometry imaging. *arXiv:1906.07102 [astro-ph]*. arXiv: 1906.07102.

Zernike, F. (1938). The concept of degree of coherence and its application to optical problems. *Physica*, 5(8):785–795.

Zheng, H., Tegmark, M., Buza, V., Dillon, J. S., Gharibyan, H., Hickish, J., Kunz, E., Liu, A., Losh, J., Lutomirski, A., Morrison, S., Narayanan, S., Perko, A., Rosner, D., Sanchez, N., Schutz, K., Tribiano, S. M., Valdez, M., Yang, H.-I., Adami, K. Z., Zelko, I., Zheng, K., Armstrong, R., Bradley, R. F., Dexter, M. R., Ewall-Wice, A., Magro, A., Matejek, M., Morgan, E., Neben, A. R., Pan, Q., Penna, R. F., Peterson, C. M., Su, M., Villasenor, J., Williams, C. L., and Zhu, Y. (2014). MITEoR: A Scalable Interferometer for Precision 21 cm Cosmology. *Monthly Notices of the Royal Astronomical Society*, 445(2):1084–1103. arXiv: 1405.5527.